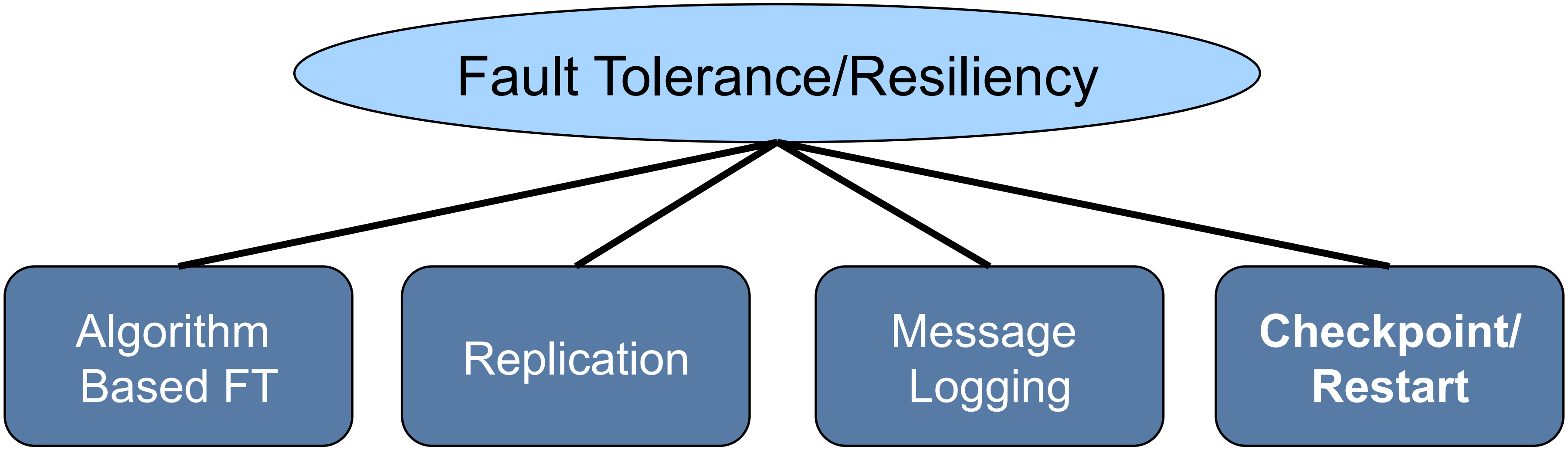
Fault Tolerance in Open MPI

Joshua Hursey

Indiana University
Open Systems Lab.
jjhursey@open-mpi.org
www.cs.indiana.edu/~jjhursey



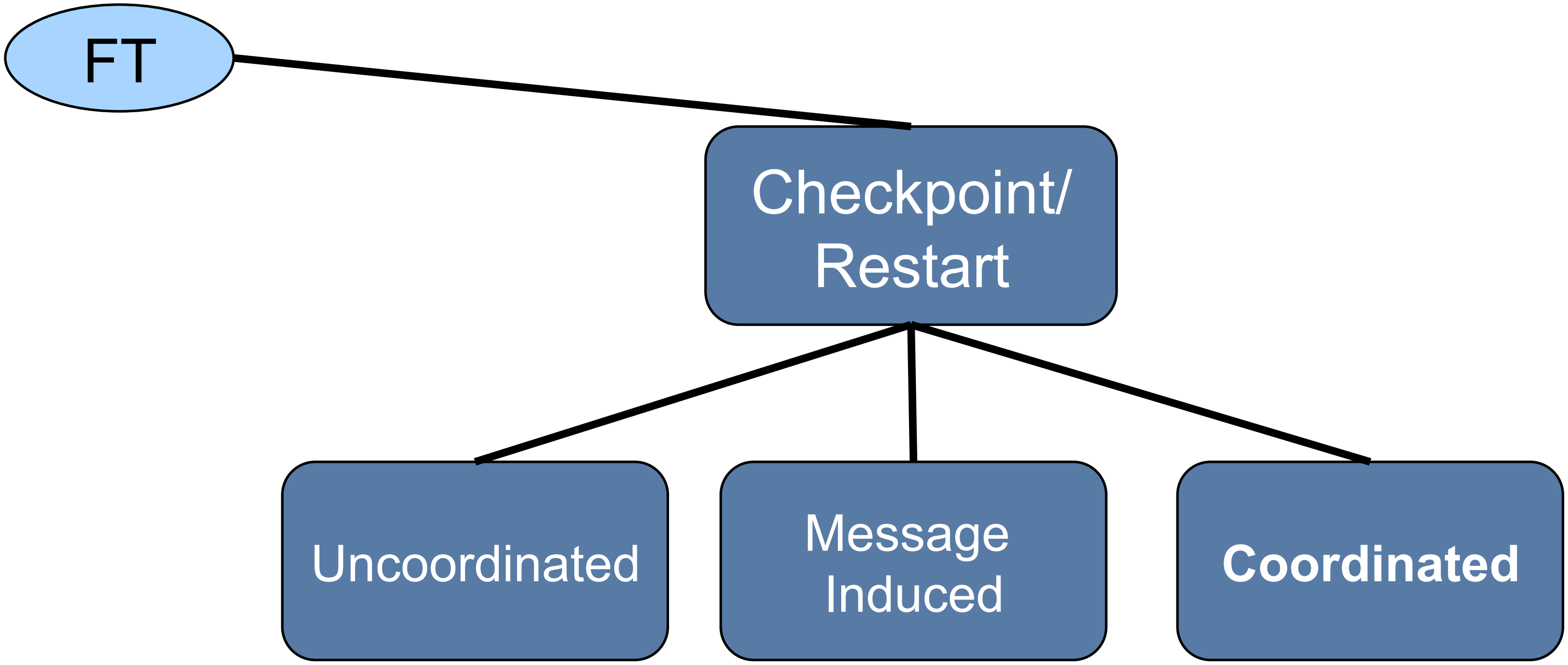
Fault Tolerance/Resiliency

Algorithm Based FT

Replication

Message Logging

Checkpoint/ Restart



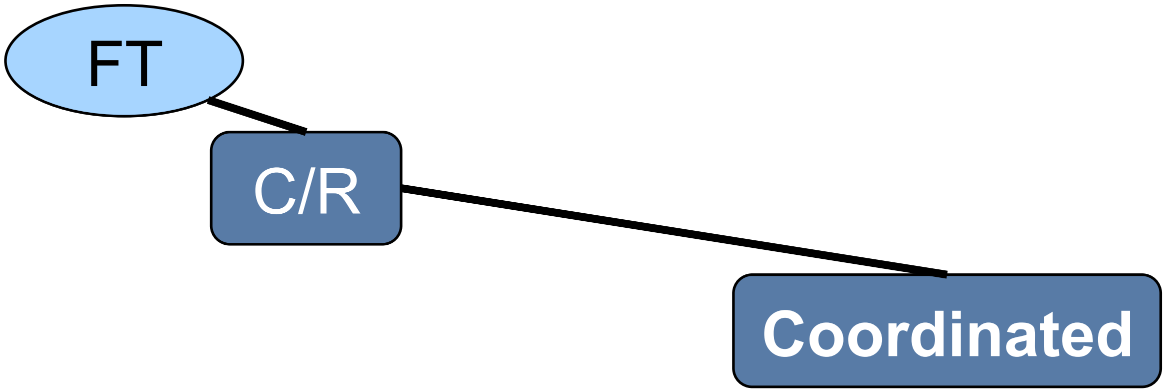
FT

Checkpoint/ Restart

Uncoordinated

Message Induced

Coordinated



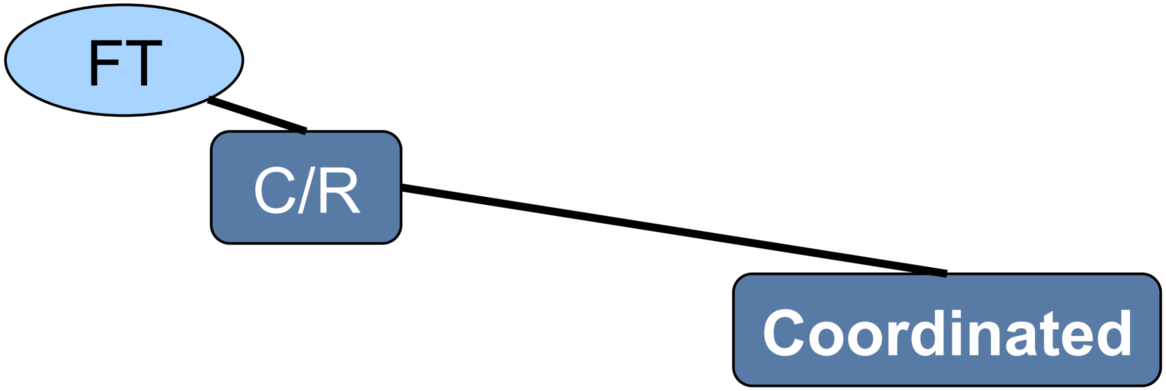
FT

C/R

Coordinated

High Level Goals

- Deliver usable features to end users
  - Don’t publish and run
- Extensible C/R research infrastructure
  - Focused development areas
  - Apples-to-apples comparisons
  - Opportunities for public release & support



FT

C/R

Coordinated

Features

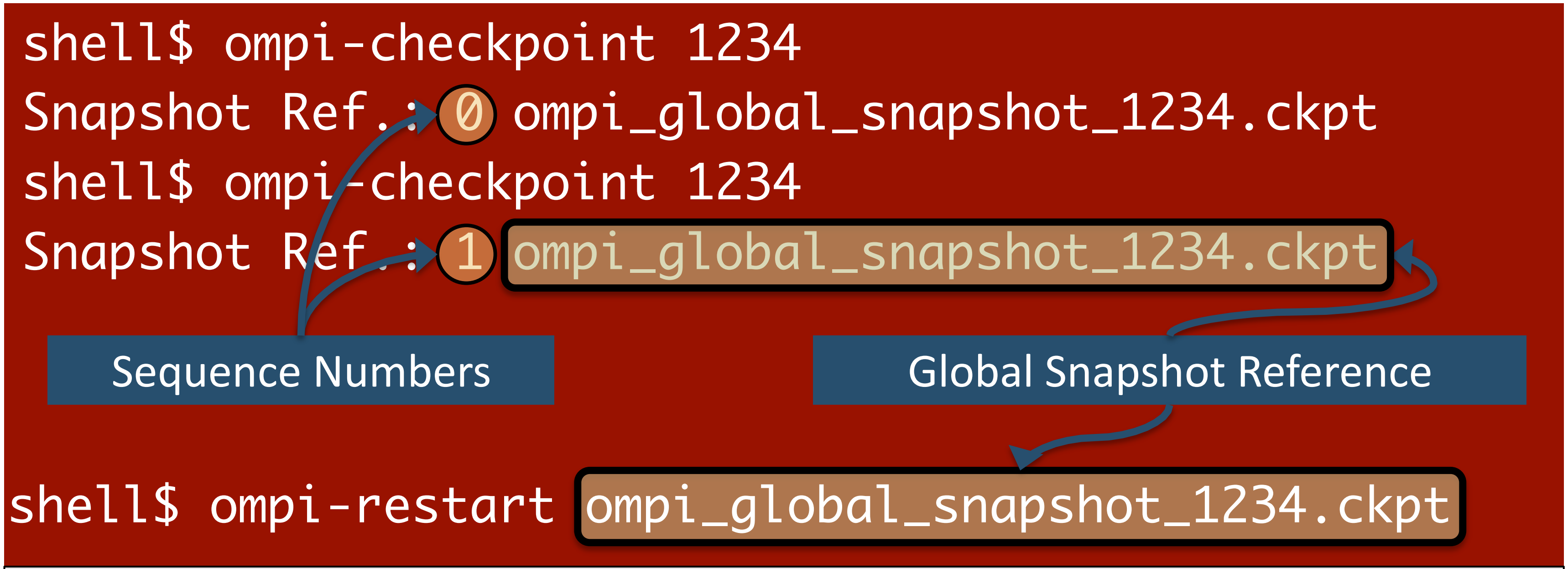
- □ Fault Tolerance
- □ Debugging
- □ Process Migration

Infrastructure

- □ Checkpoint Service
- □ Coordination Protocol
- □ Runtime Coordination
- □ File Management
- □ Internal Coordination
- □ Recovery Service
- □ *In development…*

Feature: Fault Tolerance

- Transparent, checkpoint/restart driven by:
  - System Administrator
  - Resource Manager/Scheduler
  - Application



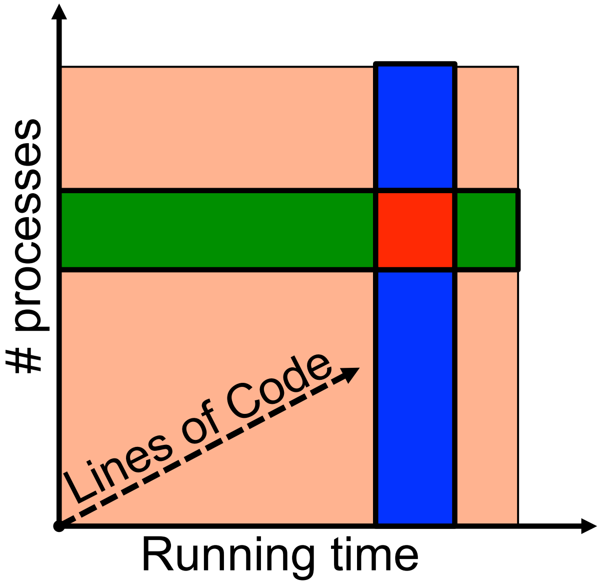
```
shell$ ompi-checkpoint 1234
Snapshot Ref.: 0 ompi_global_snapshot_1234.ckpt
shell$ ompi-checkpoint 1234
Snapshot Ref.: 1 ompi_global_snapshot_1234.ckpt

shell$ ompi-restart ompi_global_snapshot_1234.ckpt
```

Hursey, J., et. al., *The design and implementation of checkpoint/restart process fault tolerance for Open MPI*. IEEE IPDPS, 2007.

Feature: Debugging

“My program only fails after ***4 hours*** when running with ***>512 processes***.”

- Step-backward (a.k.a. reverse execution)
  - Combination of checkpoint/restart and message logging
- Specified a C/R interface for:
  - Parallel debugger,
  - C/R enabled MPI implementation,
  - Checkpoint/restart service



Hursey, J., et. al., *Checkpoint/Restart Enabled Parallel Debugging*. (under submission), 2009.

Feature: Process Migration

Transparent process migration without residual dependencies

```
shell$ ompi-migrate --off odin001 123
shell$ ompi-migrate --off odin001 --onto odin002,odin003 123
```

- Proactive Migration
  - Move processes when asked by predictor (e.g., CIFTS FTB, RAS, …)
- Cluster Management
  - Move processes when asked by end user
- Automatic Recovery
  - Rollback all processes to the last checkpoint, restart failed processes on new/spare resources.

Performance Impact

Latency

Interconnect	No C/R	With C/R	% Overhead
Ethernet (TCP)	49.92 µs	50.01 µs	0.2 %
InfiniBand	8.25 µs	8.78 µs	6.4 %
Myrinet MX	4.23 µs	4.81 µs	13.7 %
Shared Memory	1.84 µs	2.15 µs	16.8 %

Bandwidth

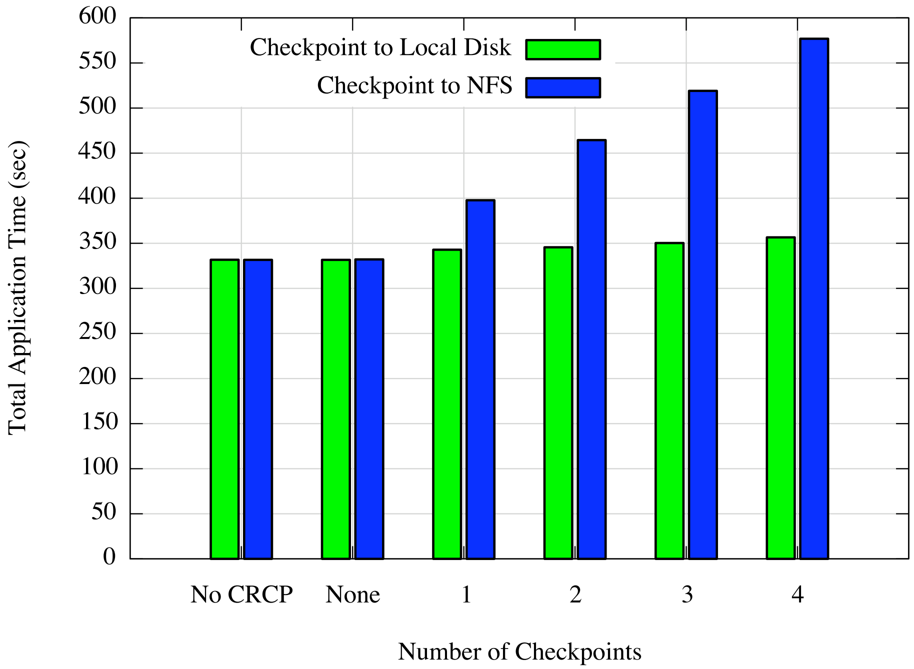
Interconnect	No C/R	With C/R	% Overhead
Ethernet (TCP)	738 Mbps	738 Mbps	0.0 %
InfiniBand	4703 Mbps	4703 Mbps	0.0 %
Myrinet MX	8000 Mbps	7985 Mbps	0.2 %
Shared Memory	5266 Mbps	5258 Mbps	0.2 %

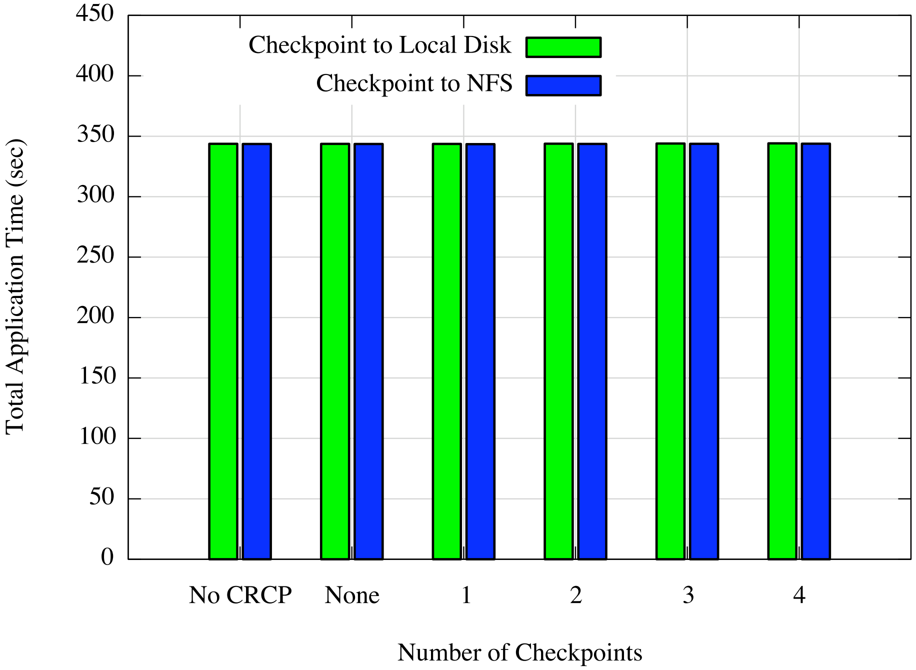
NASA Parallel Benchmarks: 0 – 0.6 %

Gromacs (DPPC): 0%

Hursey, J., et. al., *Interconnect Agnostic Checkpoint/Restart in Open MPI*. ACM HPDC, 2009.

Checkpoint Overhead



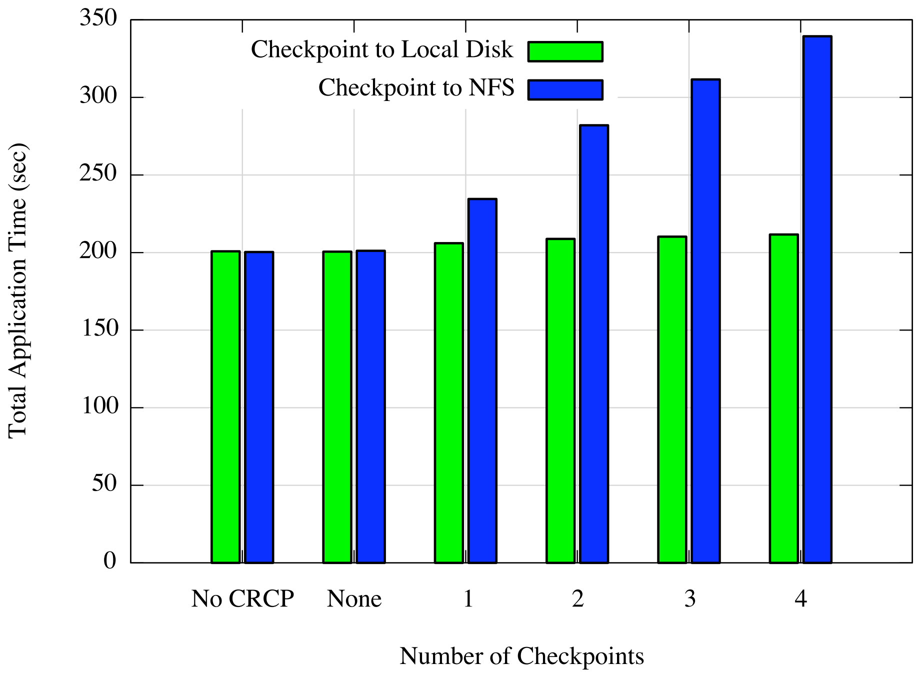


BT Class C 36 Procs
4.2 GB/120 MB

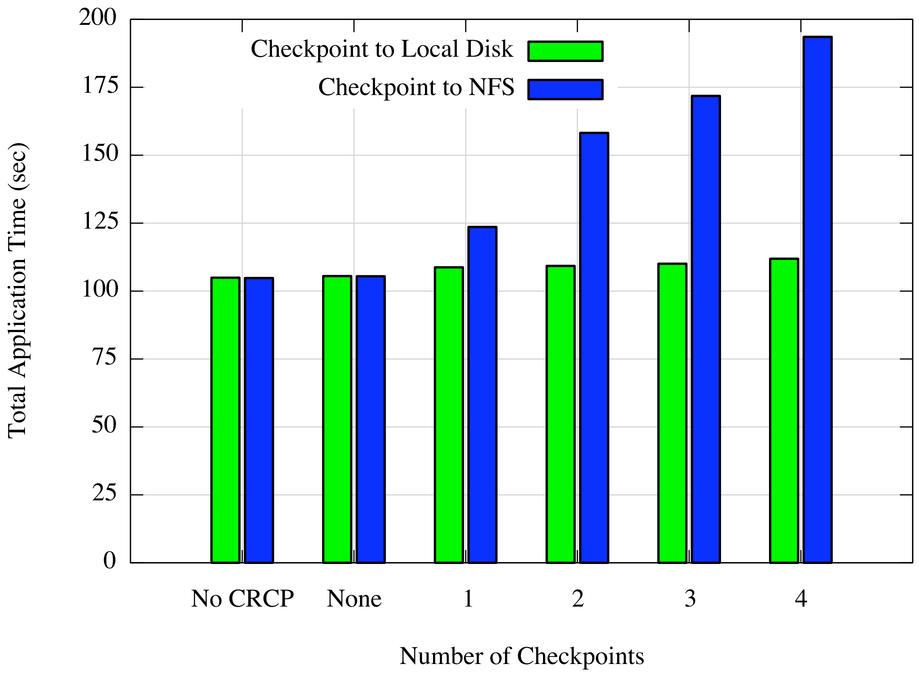
EP Class D 32 Procs
102 MB/3.2 MB

Hursey, J., et. al., *Interconnect Agnostic Checkpoint/Restart in Open MPI*. ACM HPDC, 2009.

Checkpoint Overhead



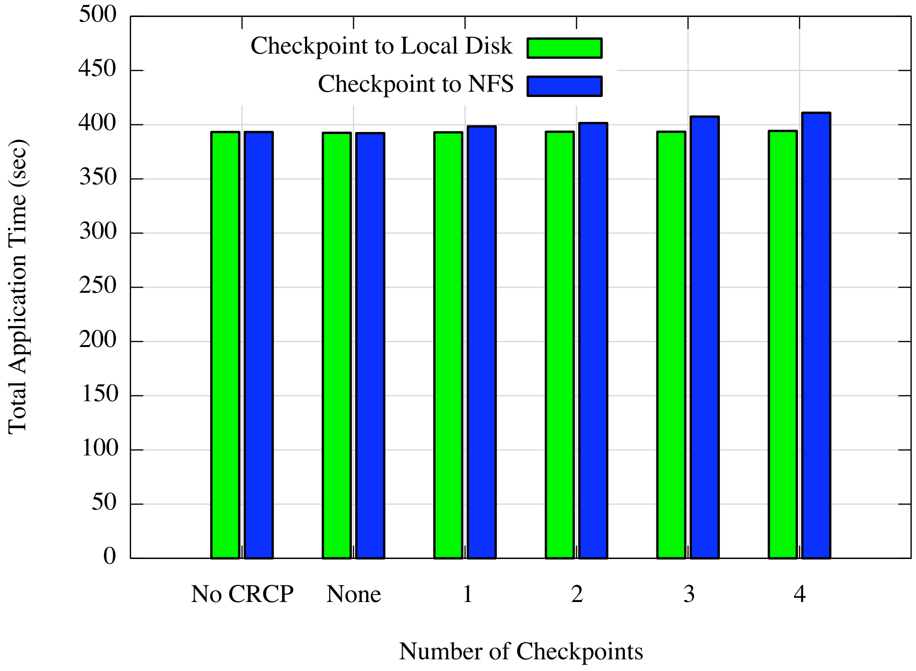
SP Class C 36 Procs
1.9 GB/54 MB

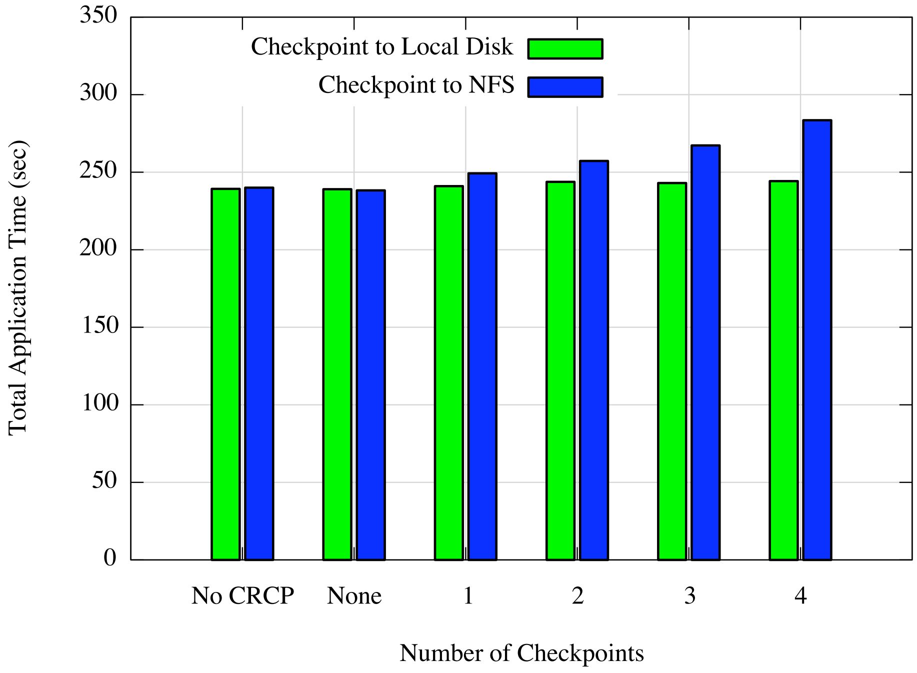


LU Class C 32 Procs
1 GB/32 MB

Hursey, J., et. al., *Interconnect Agnostic Checkpoint/Restart in Open MPI*. ACM HPDC, 2009.

Checkpoint Overhead





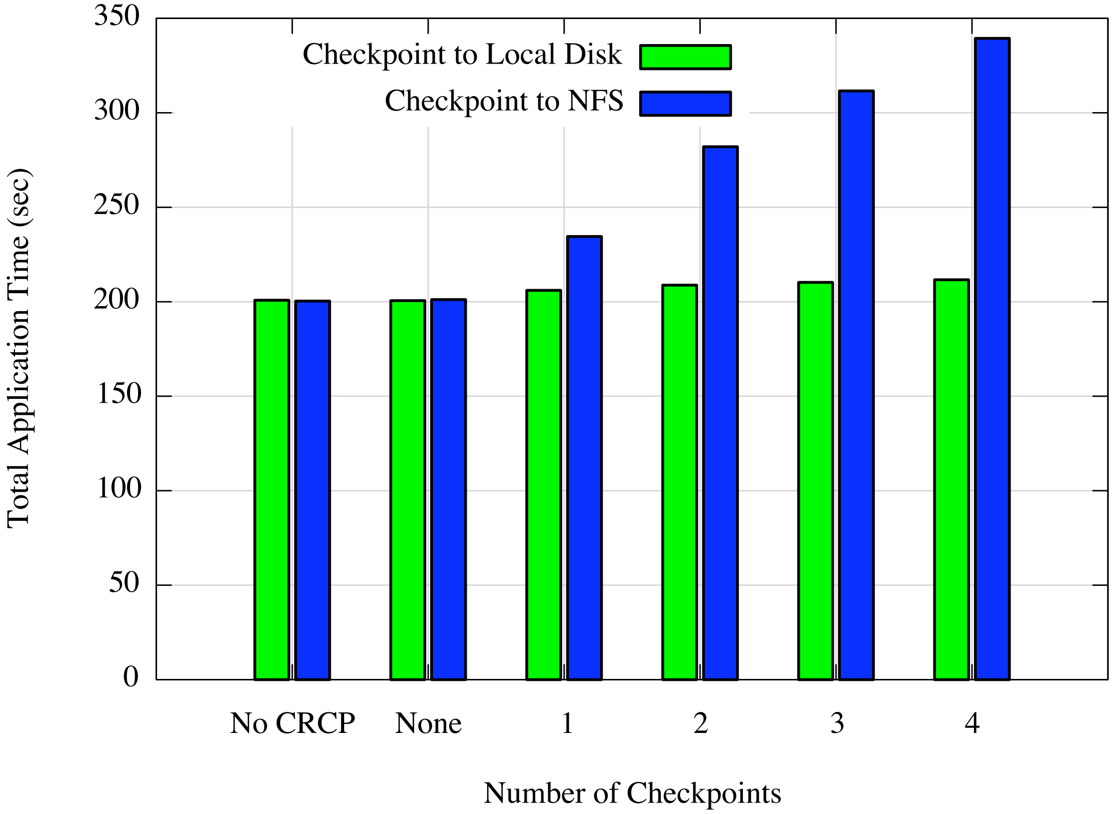
Gromacs (DPPC) 8 Procs
267 MB/33 MB

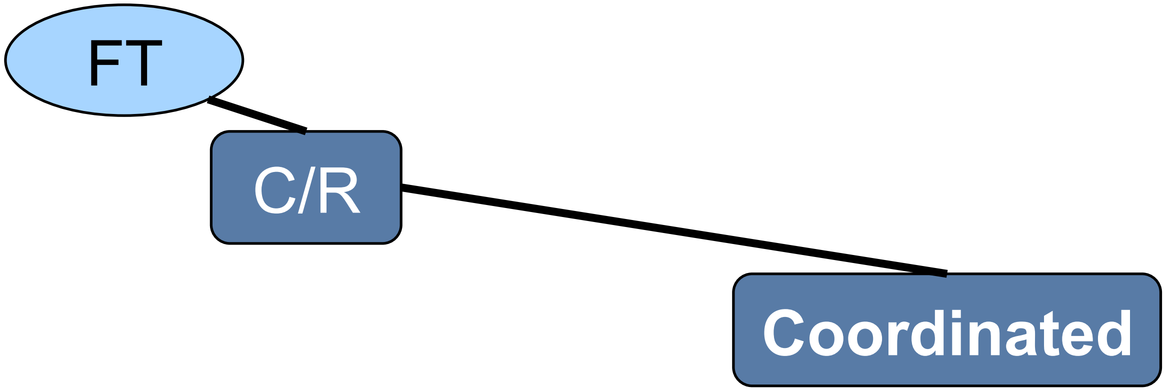
Gromacs (DPPC) 16 Procs
473 MB/30 MB

Hursey, J., et. al., *Interconnect Agnostic Checkpoint/Restart in Open MPI*. ACM HPDC, 2009.

Checkpoint Bottlenecks

98.8%	File I/O
0.7%	Modex
0.3%	Coord. Protocol
0.2%	Internal Coord.





Features

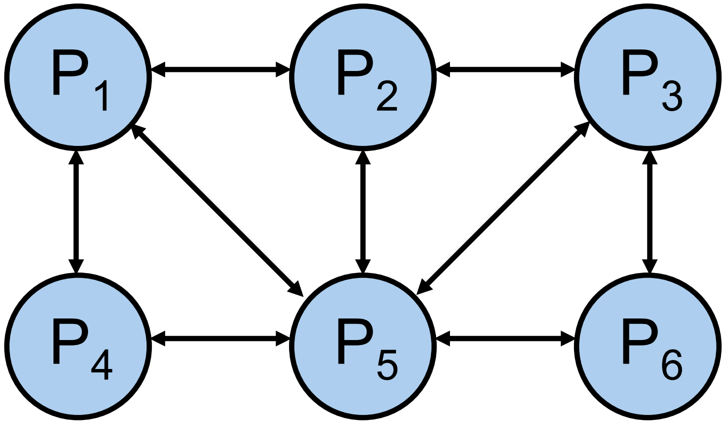
- □ Fault Tolerance
- □ Debugging
- □ Process Migration

Infrastructure

- □ Checkpoint Service
- □ Coordination Protocol
- □ Runtime Coordination
- □ File Management
- □ Internal Coordination
- □ Recovery Service
- □ *In development…*

Distributed Snapshots

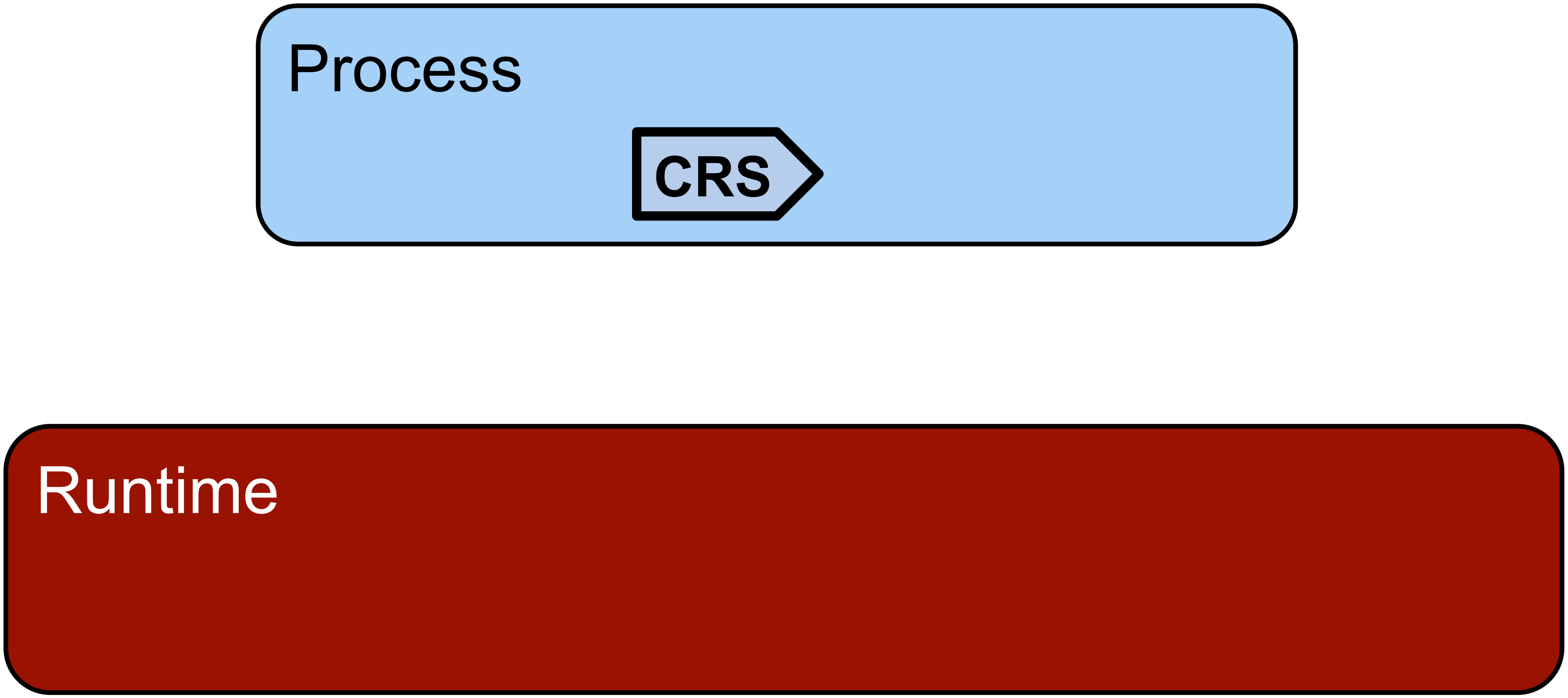
The global state of a distributed system is defined as the *state of all processes* and *all connected channels* in the system.



6 processes + 9 channels

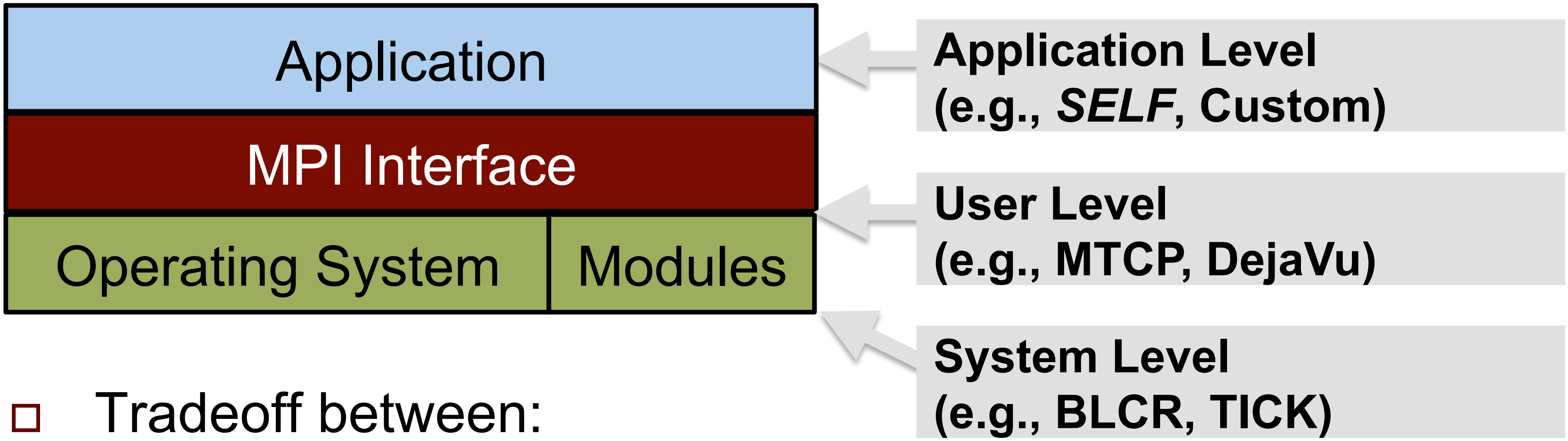
Chandy, K., Lamport, L. *Distributed snapshots: Determining global states of distributed systems*. ACM Transactions on Computer Systems (TOCS), 1985

C/R Infrastructure in Open MPI



Checkpoint/Restart Service (CRS)

Capture the state of a single process



Application Level (e.g., *SELF*, Custom)

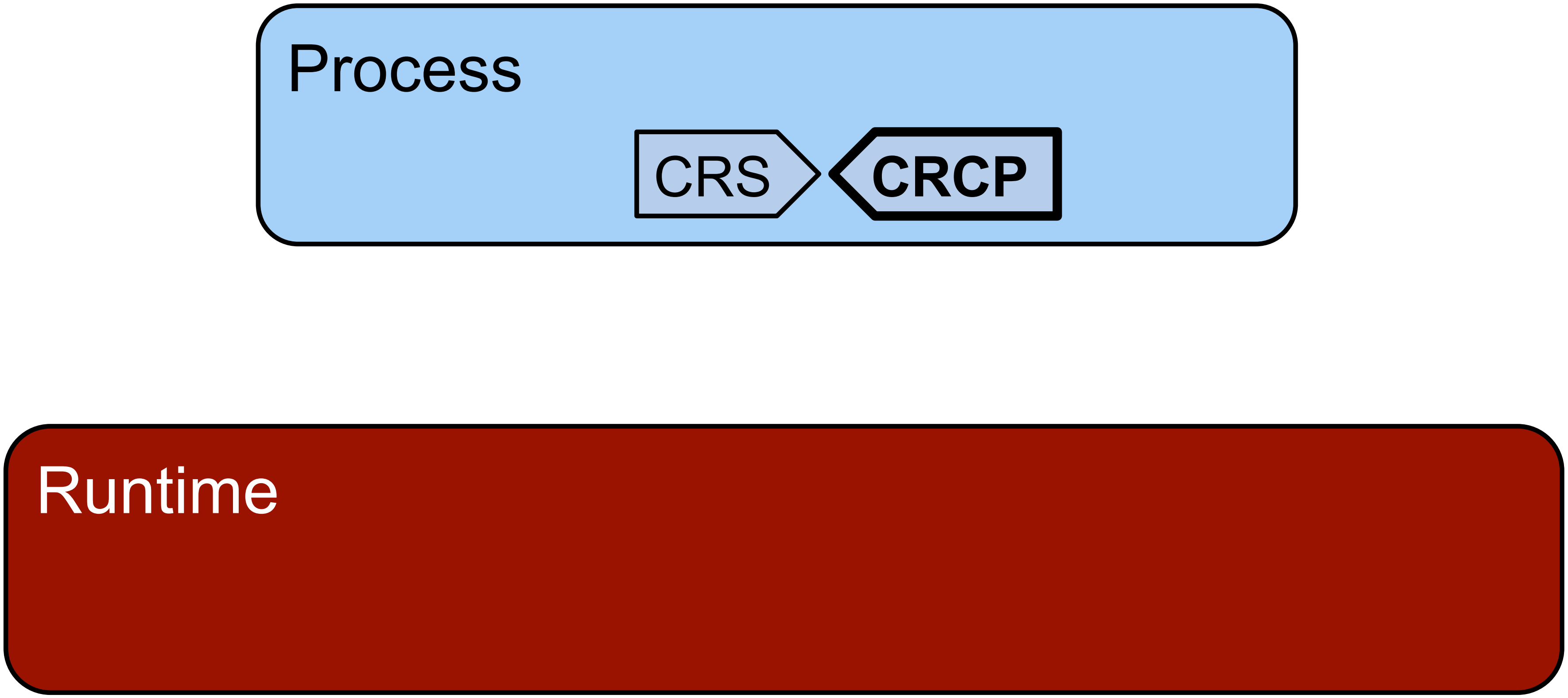
User Level (e.g., MTCP, DejaVu)

System Level (e.g., BLCR, TICK)

- Tradeoff between:
  - Transparency
  - Performance
  - Portability
- API and/or callbacks required for MPI support

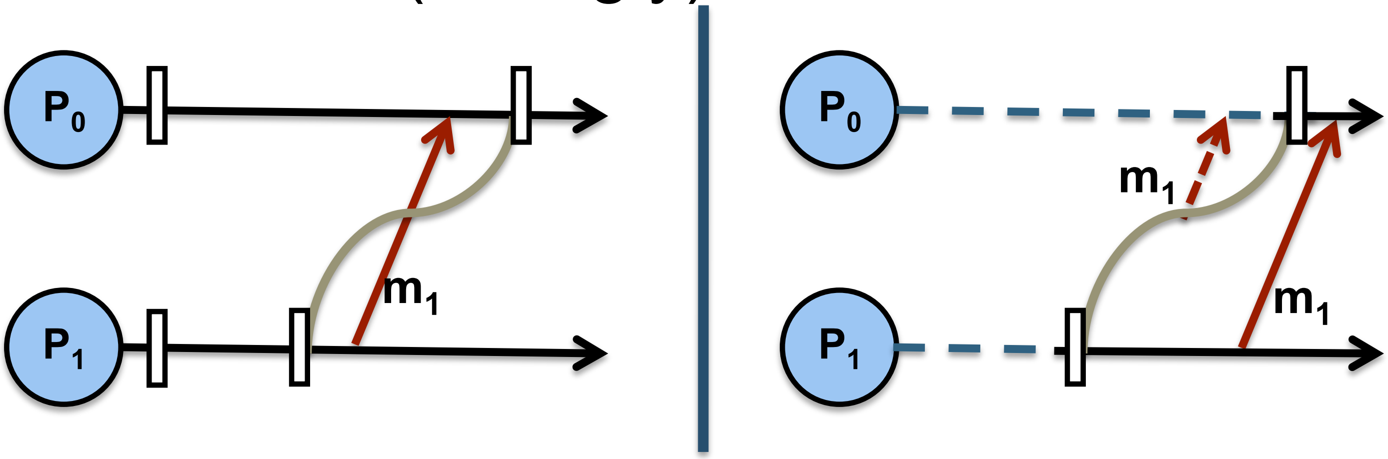
Hursey, J., et. al., *A Checkpoint and Restart Service Specification for Open MPI*. IU Tech. Report TR635, 2006.

C/R Infrastructure in Open MPI



Message Coordination Protocol

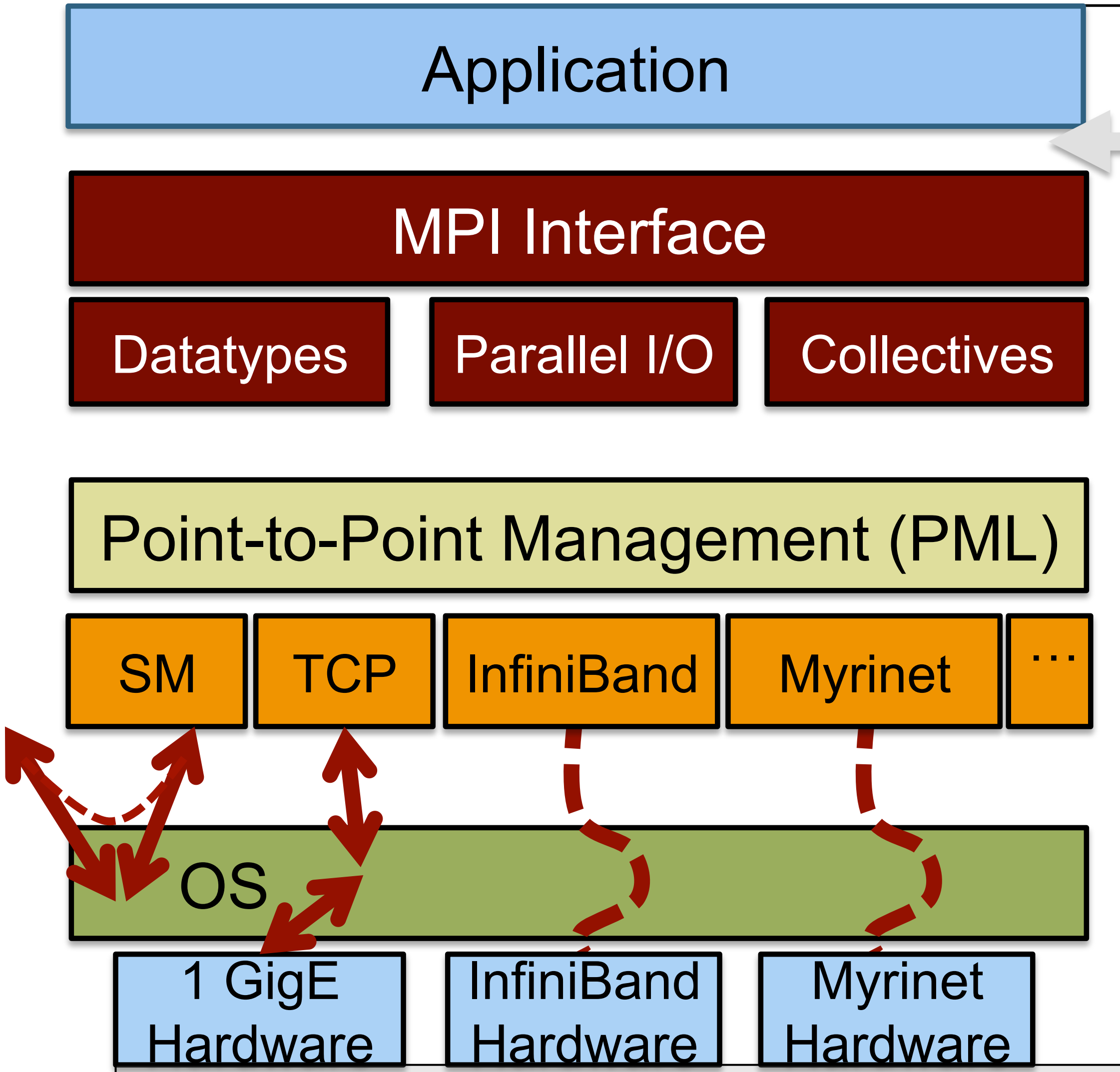
Capture the state of all connected channels.
Find a (strongly) consistent state.



- Common Coordination Algorithms
  - Chandy/Lamport's Distributed Snapshots
  - CoCheck's Ready Message
  - LAM/MPI's Bookmark Exchange

Hursey, J., et. al., *The design and implementation of checkpoint/restart process fault tolerance for Open MPI*. IEEE IPDPS, 2007.

Coordination Protocol Integration



Application

MPI Interface

Datatypes

Parallel I/O

Collectives

Point-to-Point Management (PML)

SM

TCP

InfiniBand

Myrinet

…

OS

1 GigE Hardware

InfiniBand Hardware

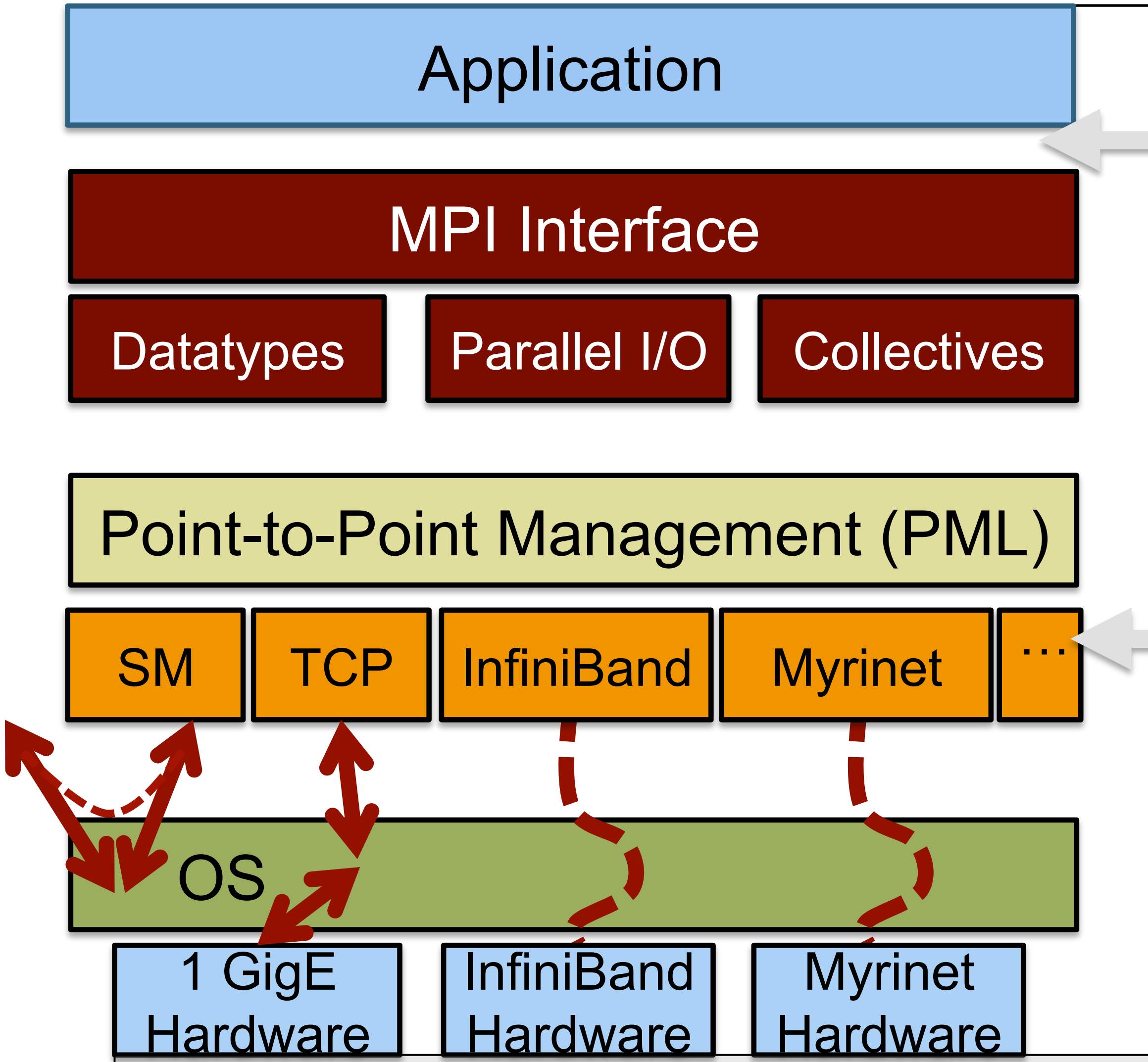
Myrinet Hardware

MPI Virtualization

- Complex
  - ~300 functions
- Flexible
  - Any Network & MPI

Hursey, J., et. al., *Interconnect Agnostic Checkpoint/Restart in Open MPI*. ACM HPDC, 2009.

Coordination Protocol Integration



Application

MPI Interface

Datatypes

Parallel I/O

Collectives

Point-to-Point Management (PML)

SM

TCP

InfiniBand

Myrinet

…

OS

1 GigE Hardware

InfiniBand Hardware

Myrinet Hardware

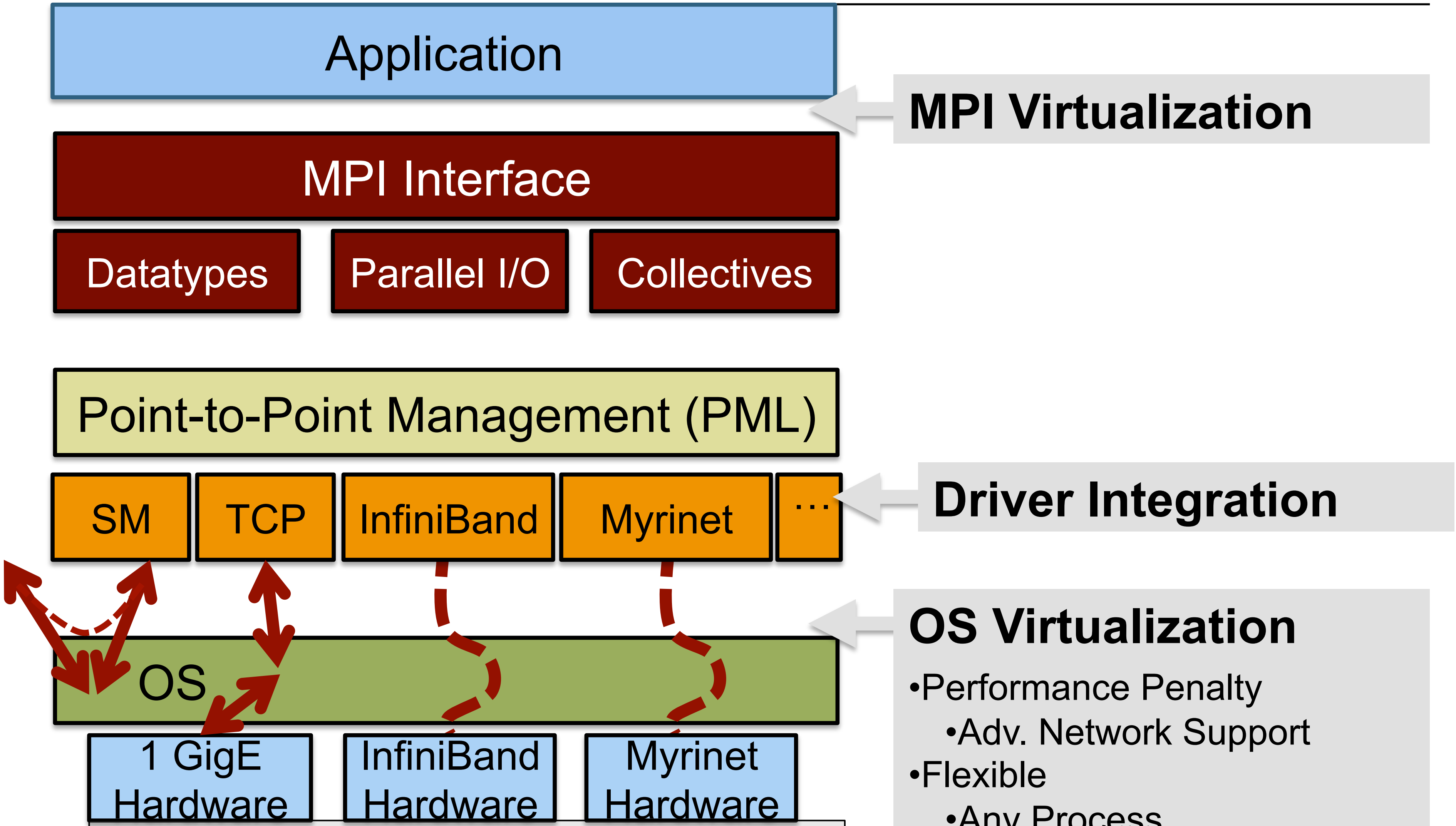
MPI Virtualization

Driver Integration

- Relatively Simple
  - Track bytes
- Flexibility Issues
  - Restart with same network
  - Muddled Coordination Alg.

Hursey, J., et. al., *Interconnect Agnostic Checkpoint/Restart in Open MPI*. ACM HPDC, 2009.

Coordination Protocol Integration



MPI Virtualization

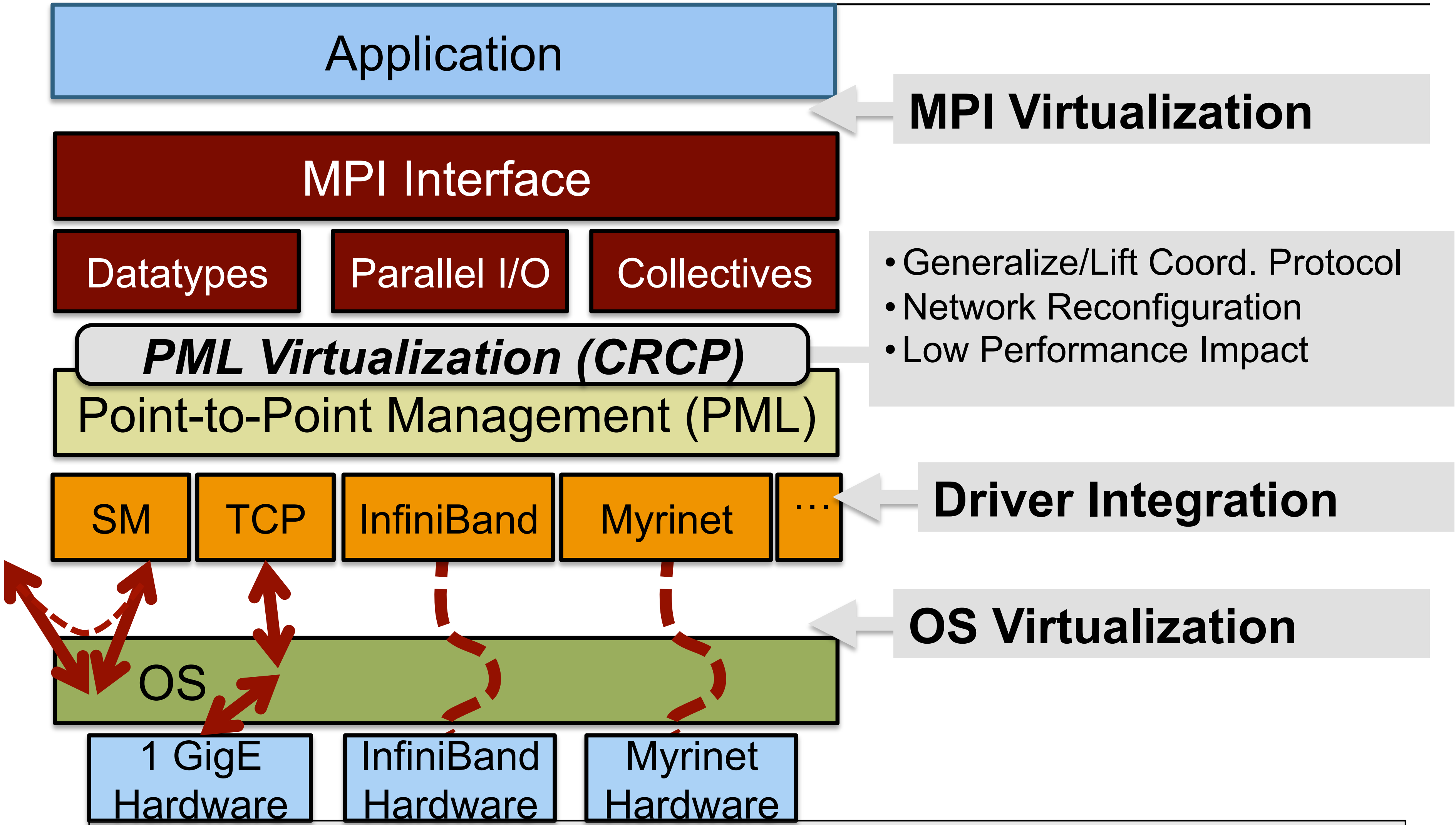
Driver Integration

OS Virtualization

- Performance Penalty
  - Adv. Network Support
- Flexible
  - Any Process

Hursey, J., et. al., *Interconnect Agnostic Checkpoint/Restart in Open MPI*. ACM HPDC, 2009.

Coordination Protocol Integration



Application

MPI Virtualization

MPI Interface

Datatypes

Parallel I/O

Collectives

PML Virtualization (CRCP)

Point-to-Point Management (PML)

- Generalize/Lift Coord. Protocol
- Network Reconfiguration
- Low Performance Impact

SM

TCP

InfiniBand

Myrinet

…

Driver Integration

OS

OS Virtualization

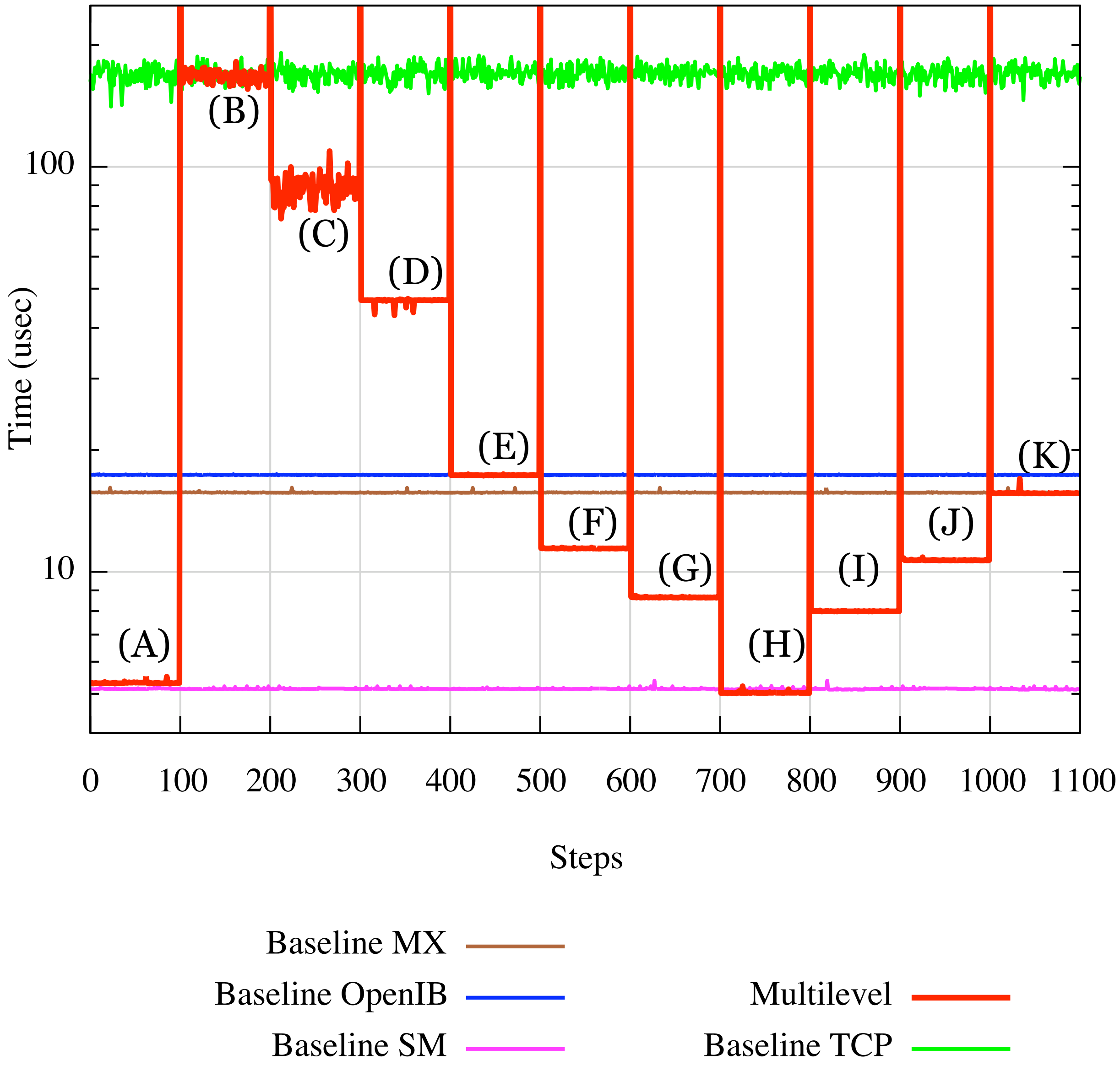
1 GigE Hardware

InfiniBand Hardware

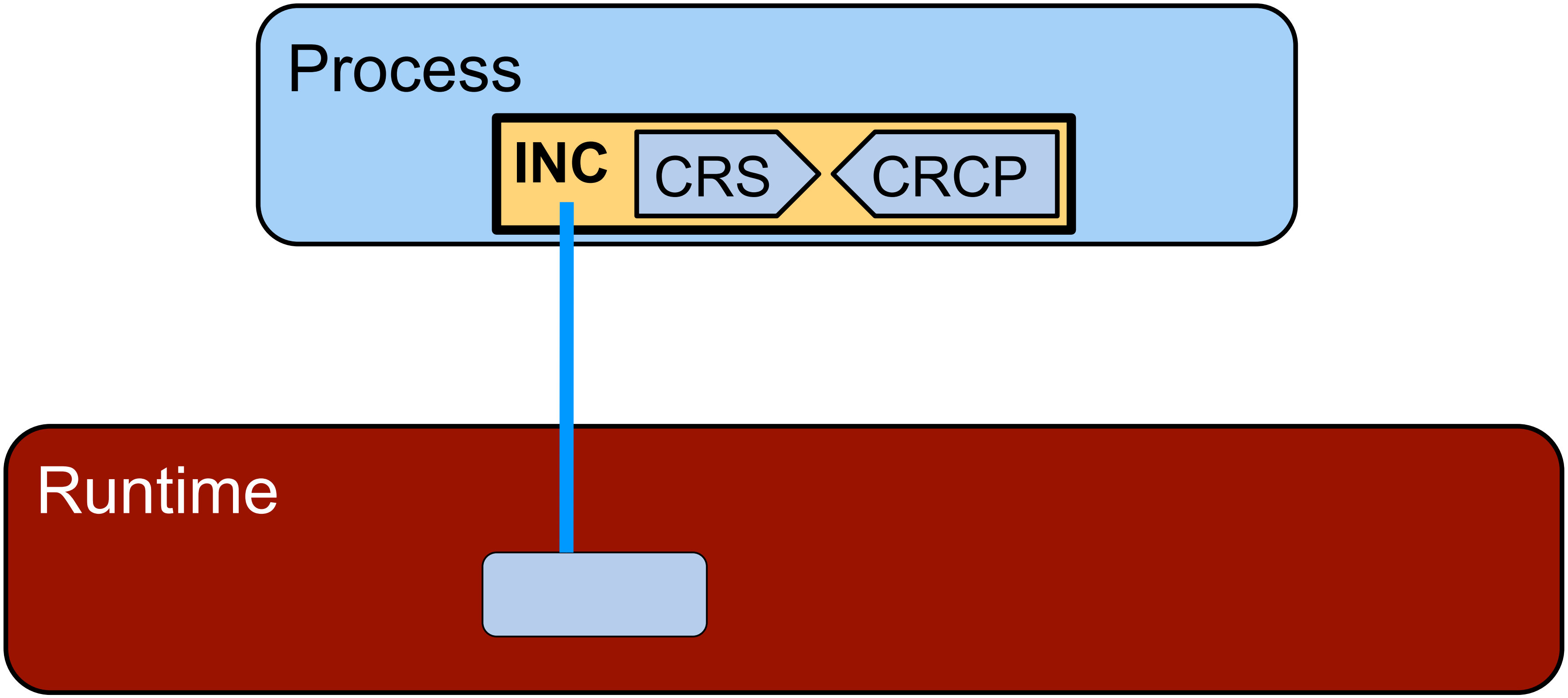
Myrinet Hardware

Hursey, J., et. al., *Interconnect Agnostic Checkpoint/Restart in Open MPI*. ACM HPDC, 2009.

Network Reconfiguration

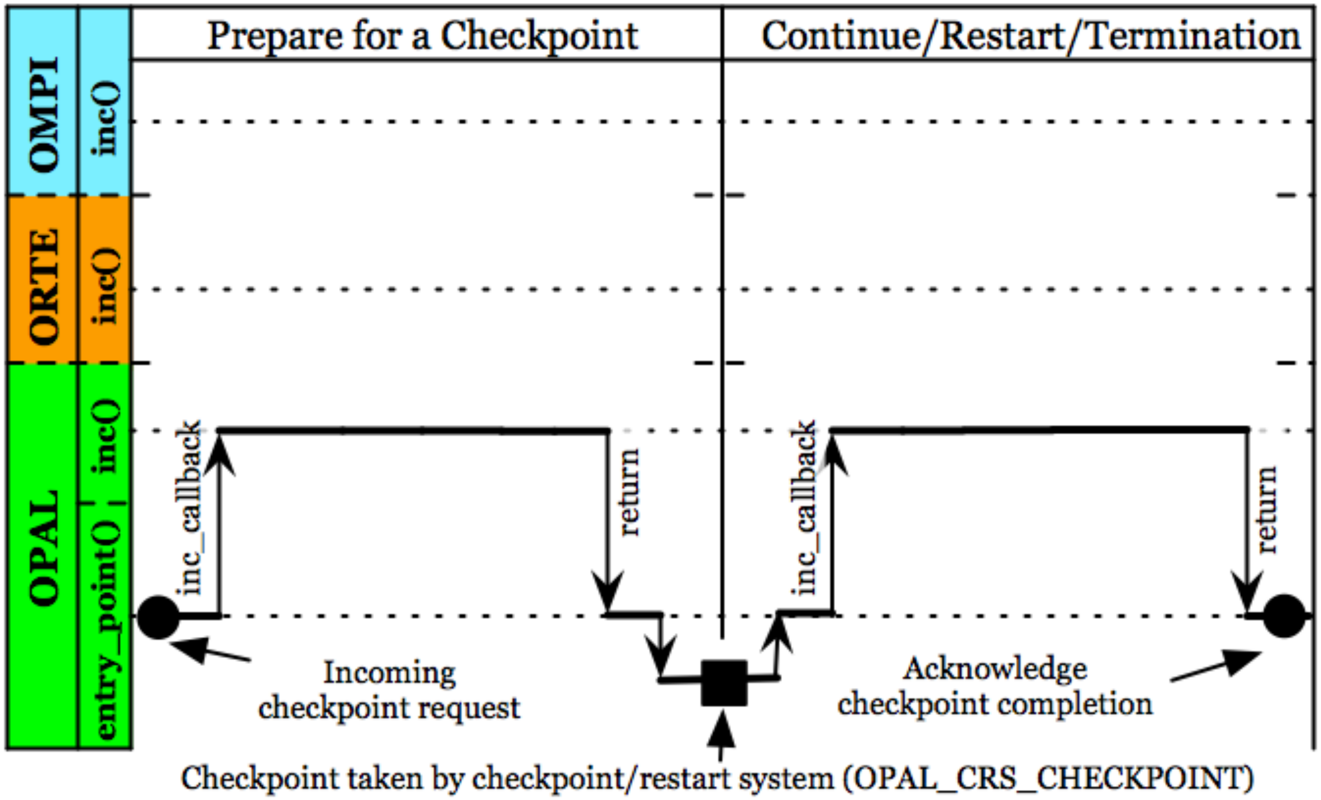


C/R Infrastructure in Open MPI



Internal Coordination (INC)

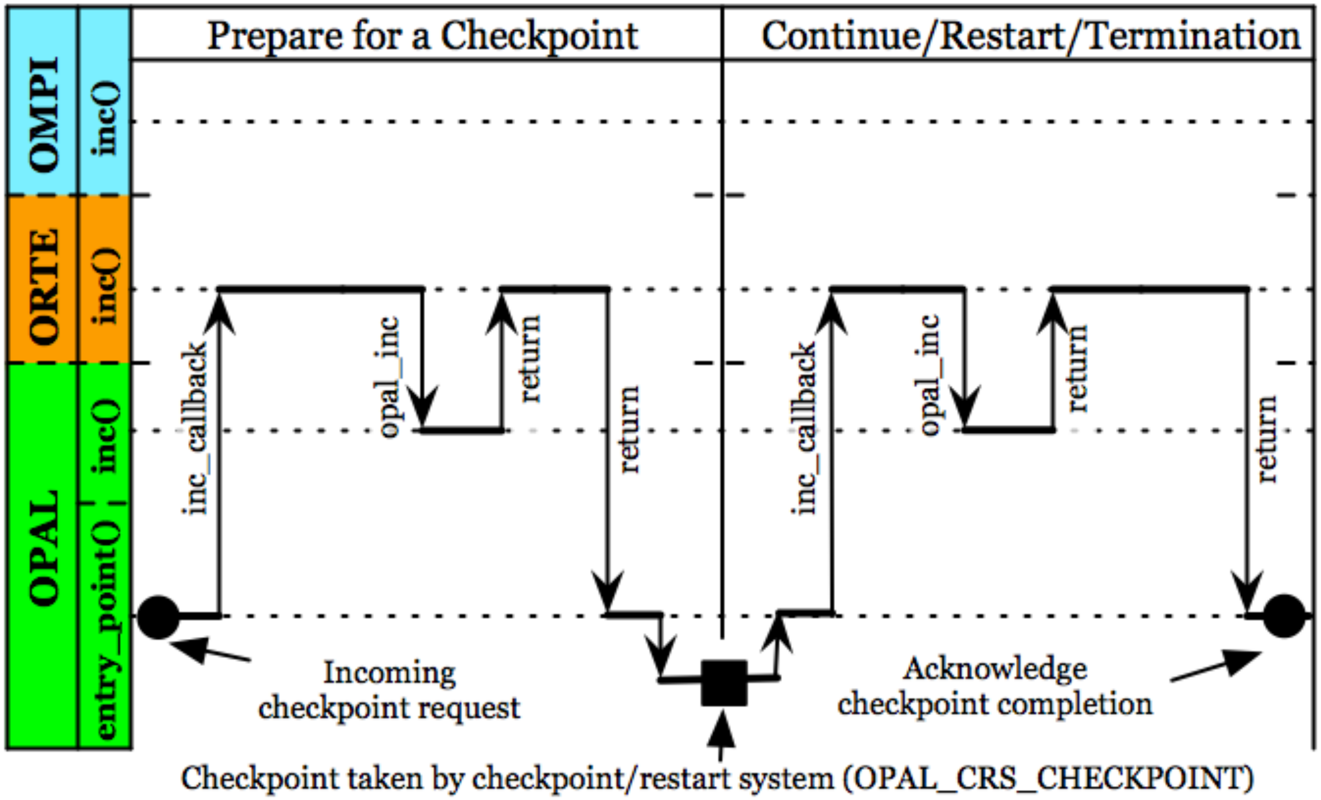
Intra-process coordination of notifications to all layers and frameworks in Open MPI


Prepare for a Checkpoint
Continue/Restart/Termination
OMPI
inc()
ORTE
inc()
OPAL
entry_point()
inc()
inc_callback
return
inc_callback
return
Incoming checkpoint request
Acknowledge checkpoint completion
Checkpoint taken by checkpoint/restart system (OPAL_CRS_CHECKPOINT)

Hursey, J., et. al., *The design and implementation of checkpoint/restart process fault tolerance for Open MPI*. IEEE IPDPS, 2007.

Internal Coordination (INC)

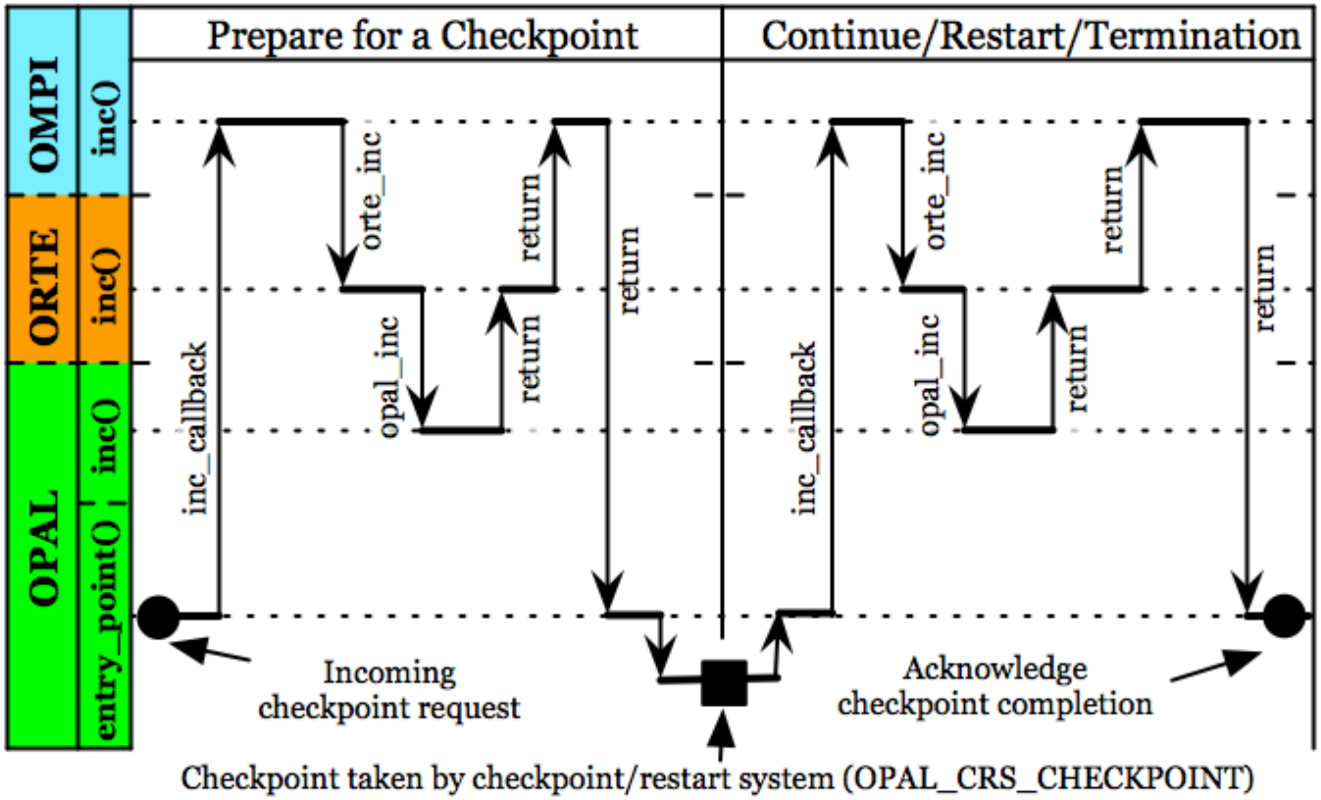
Intra-process coordination of notifications to all layers and frameworks in Open MPI


Prepare for a Checkpoint
Continue/Restart/Termination
OMPI
inc()
ORTE
inc()
OPAL
inc()
entry_point()
inc_callback
opal_inc
return
return
inc_callback
opal_inc
return
return
Incoming checkpoint request
Acknowledge checkpoint completion
Checkpoint taken by checkpoint/restart system (OPAL_CRS_CHECKPOINT)

Hursey, J., et. al., *The design and implementation of checkpoint/restart process fault tolerance for Open MPI*. IEEE IPDPS, 2007.

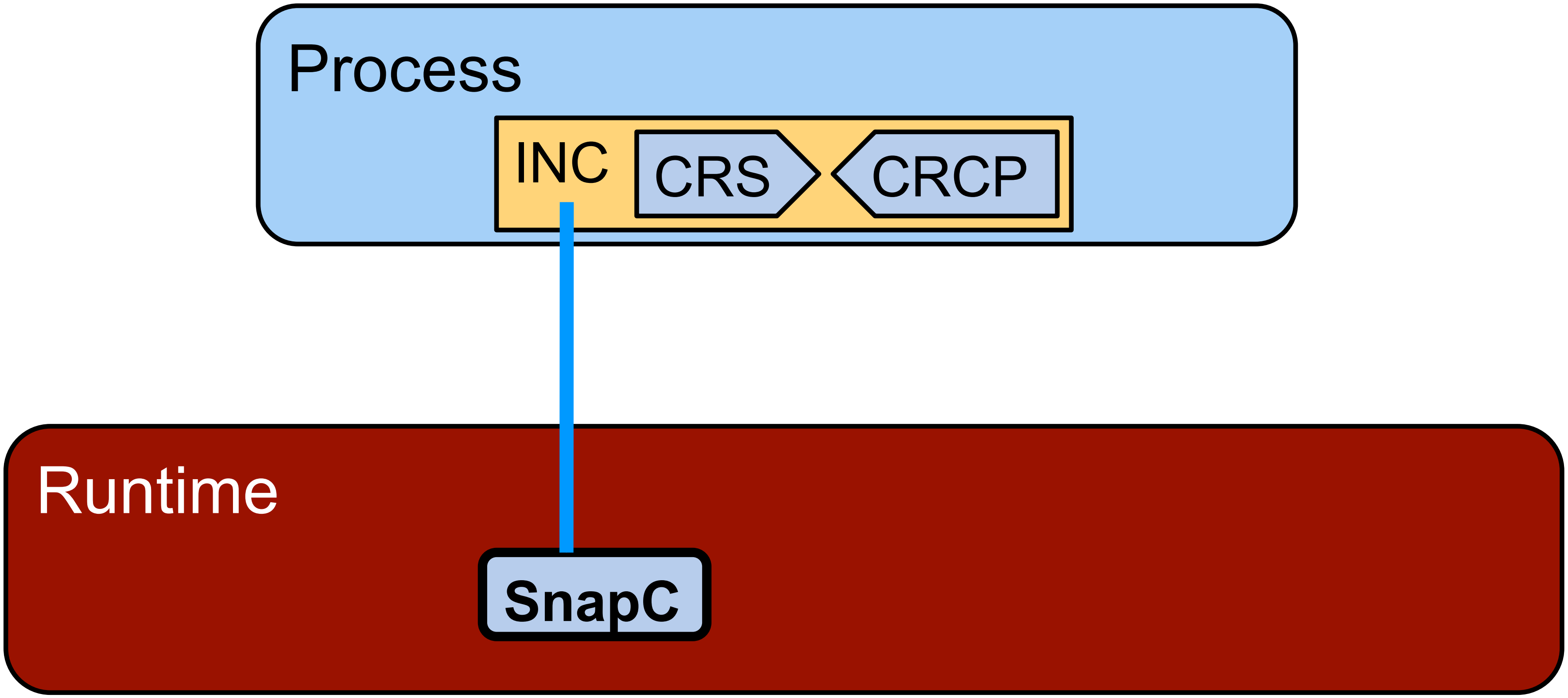
Internal Coordination (INC)

Intra-process coordination of notifications to all layers and frameworks in Open MPI



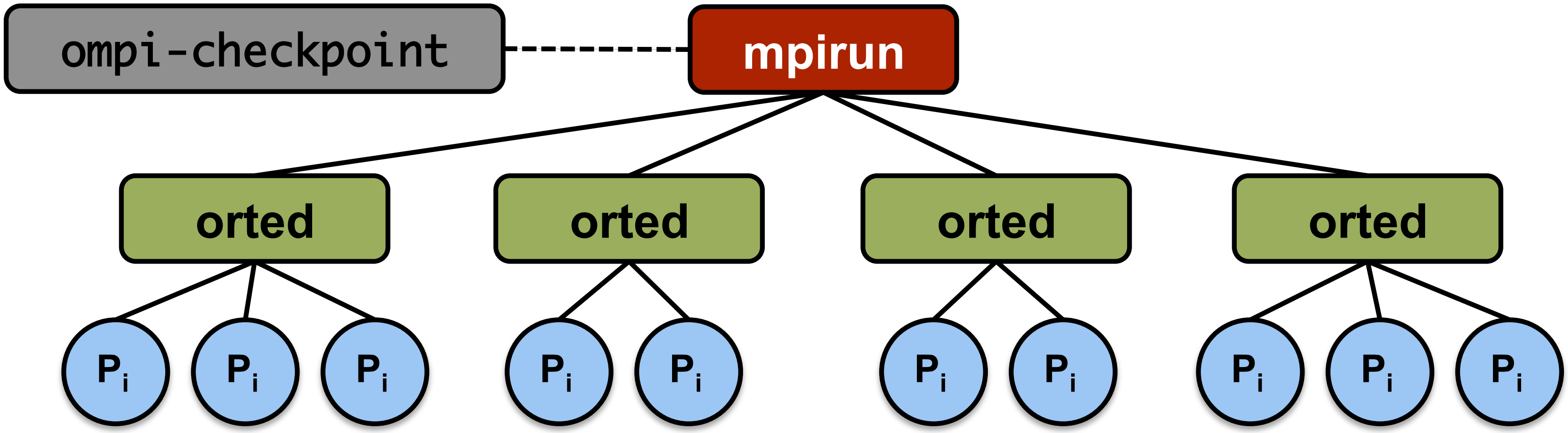
Hursey, J., et. al., *The design and implementation of checkpoint/restart process fault tolerance for Open MPI*. IEEE IPDPS, 2007.

C/R Infrastructure in Open MPI



Runtime Coordination (SnapC)

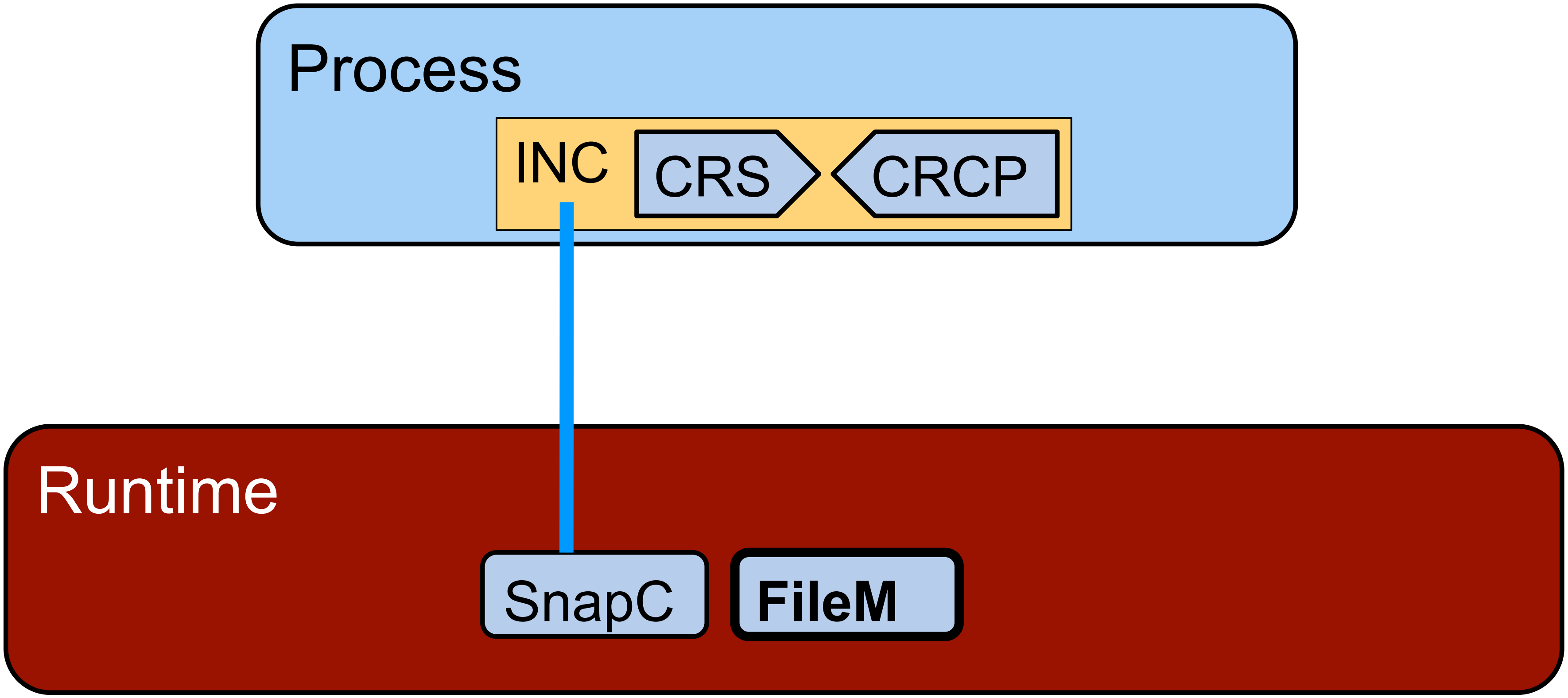
Coordinate all checkpoint related activities in ORTE, and interact with command line tools



1. Initiate the per process local checkpoint operation
2. Monitor the progress of the checkpoint operation
3. Aggregate the local snapshots into a global snapshot
4. Preserve global snapshot on stable storage

Hursey, J., et. al., *The design and implementation of checkpoint/restart process fault tolerance for Open MPI*. IEEE IPDPS, 2007.

C/R Infrastructure in Open MPI



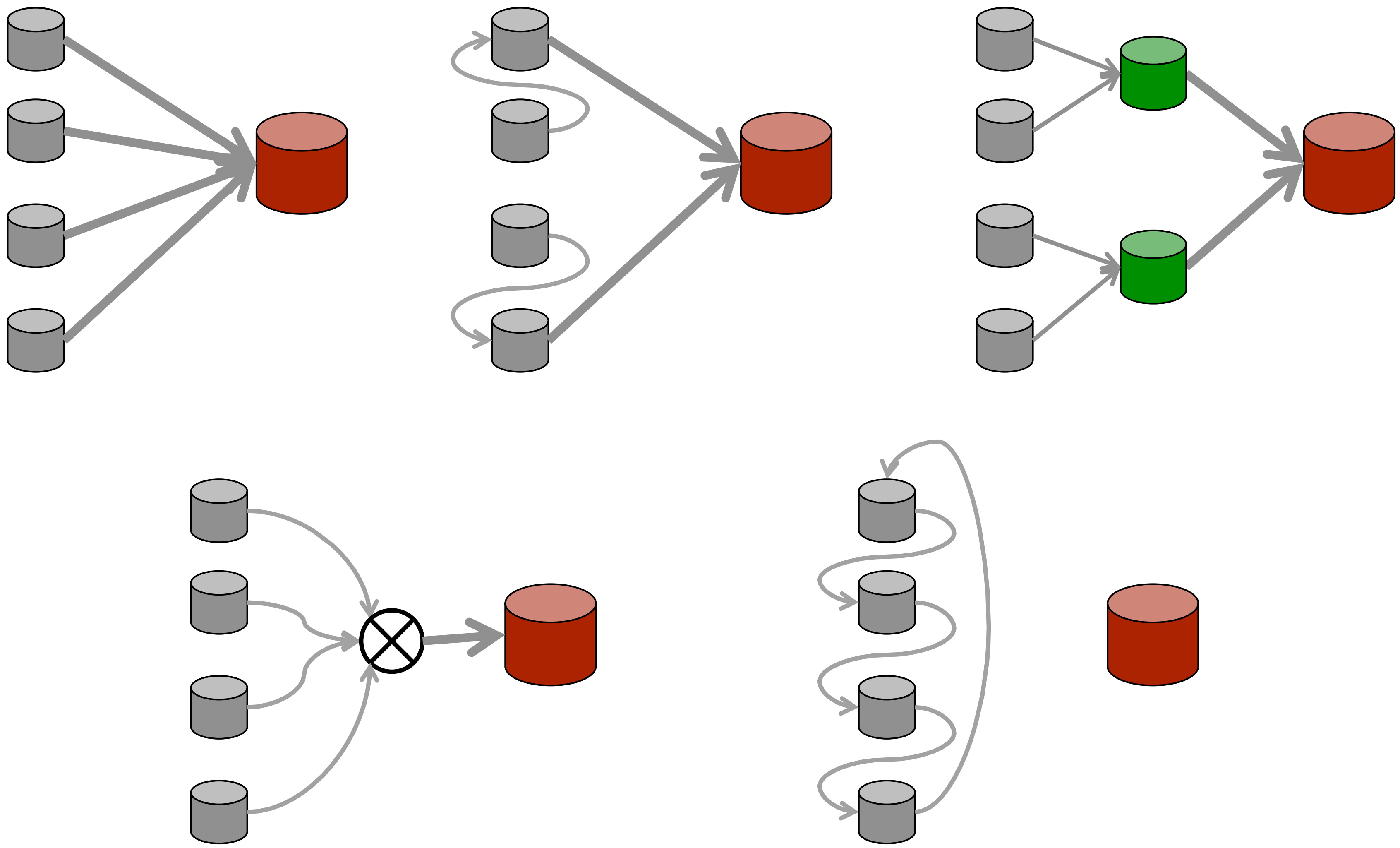
File Management (FileM)

Management the movement of files from one file system to another

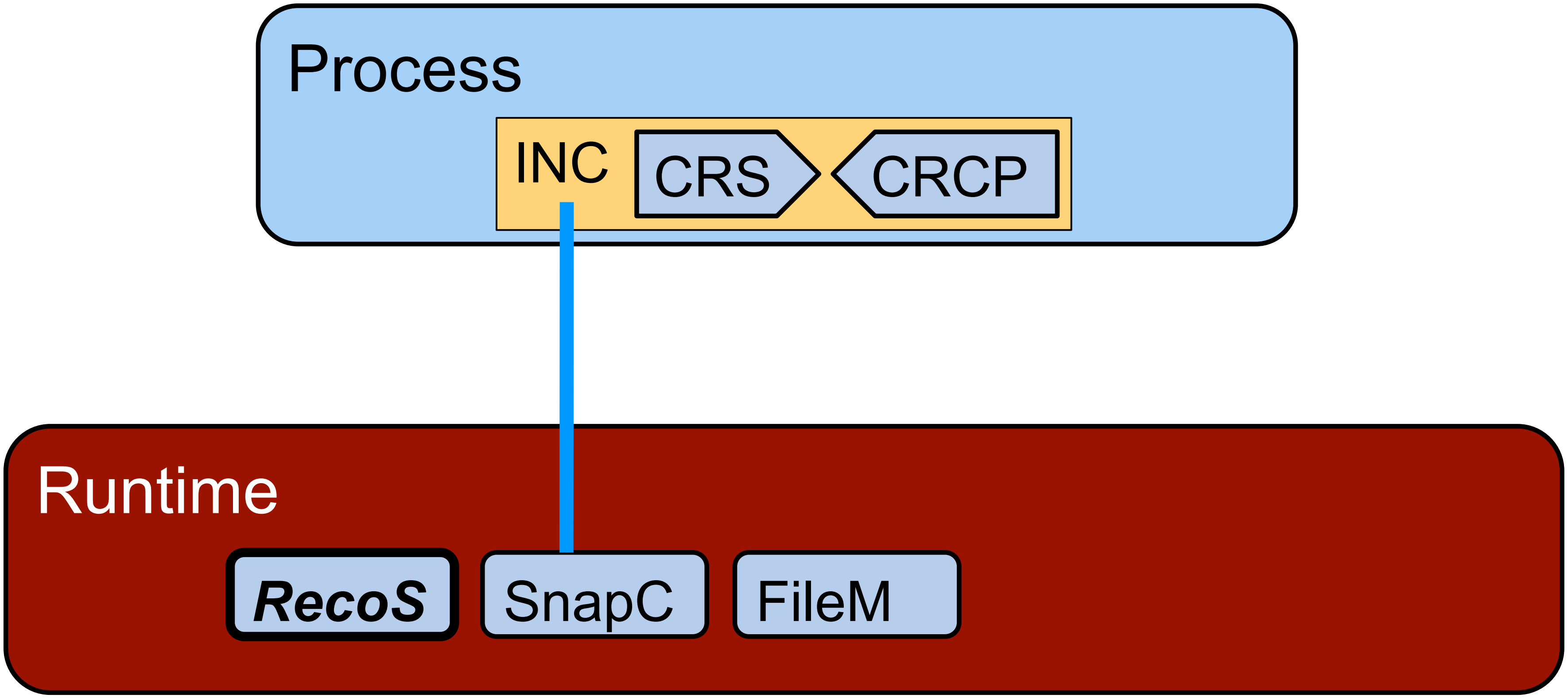
- **Stable Storage:**
  - *Any storage device that survives the maximum number of expected faults in a system*
- Interface:
  - Get() – Gather to stable storage
  - Put() – Broadcast to local storage
  - Remove() – Cleanup temporary files

Hursey, J., et. al., *The design and implementation of checkpoint/restart process fault tolerance for Open MPI*. IEEE IPDPS, 2007.

File Management (FileM)



C/R Infrastructure in Open MPI

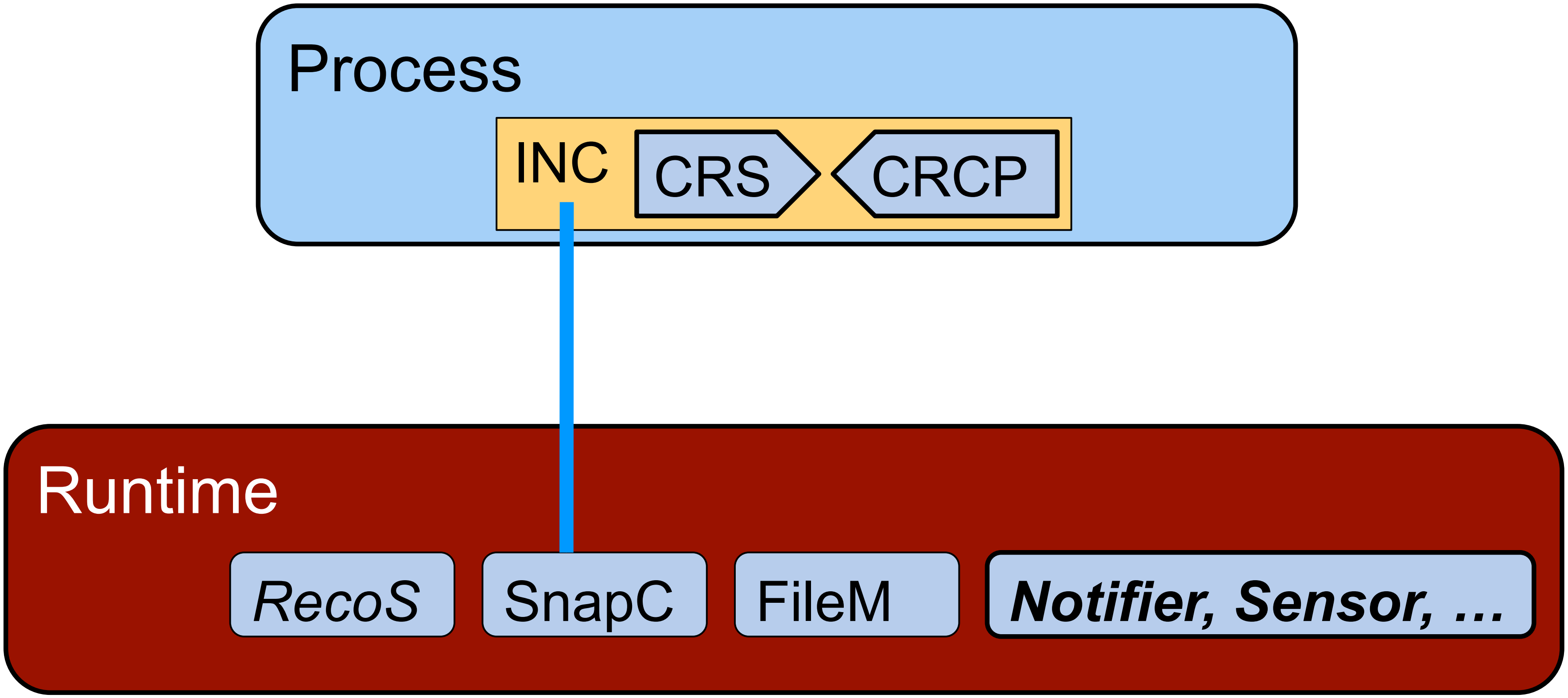


Recovery Service (RecoS)

Policy enforcement for runtime fault recovery and preventative actions

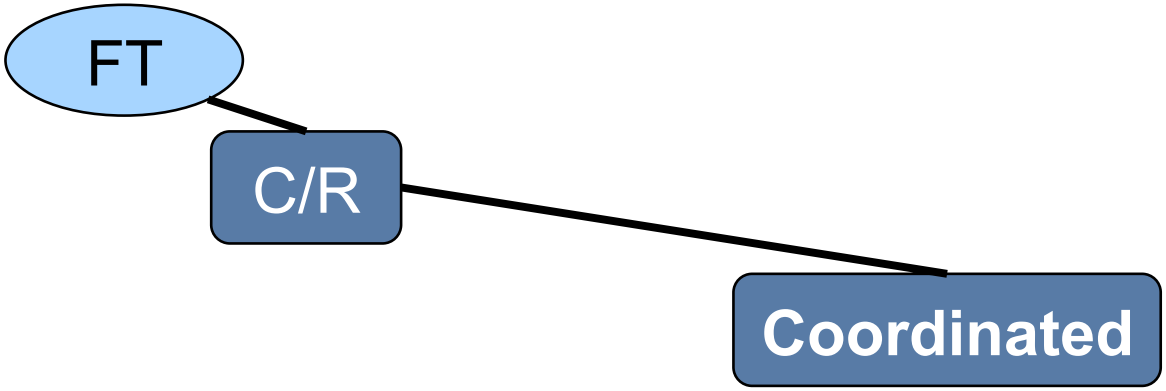
- Policy variations:
  - **Abort**:
    Terminate job
  - **Ignore**:
    Stabilize and run without the failed process
  - **Migrate**:
    Preventatively move processes between resources
  - **Restart**:
    Automatically restart from the last checkpoint
- Can be used to support MPI layer policies optionally expressed by an application

C/R Infrastructure in Open MPI



Services in Development

- Event Monitoring
  - CIFTS FTB, IPMI, and others
- Fault Prediction
  - Experimenting with models using:
    - Old logs (historical perspective)
    - Current logs (recent history)
    - Hardware sensors (present environment)
- Fault Detection
  - Current is simple heartbeat mechanism
  - Looking at more scalable protocols

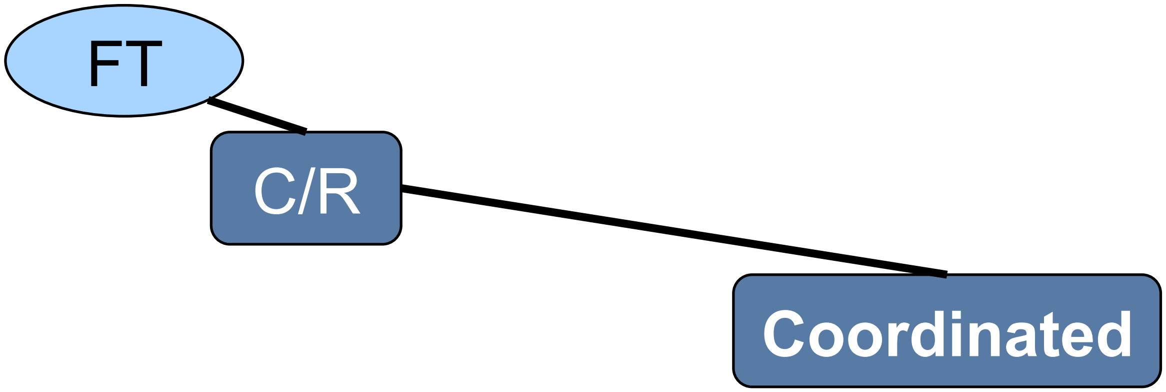


Features

- Fault Tolerance
- Debugging
- Process Migration

Infrastructure

- Checkpoint Service
- Coordination Protocol
- Runtime Coordination
- File Management
- Internal Coordination
- Recovery Service
- *In development…*



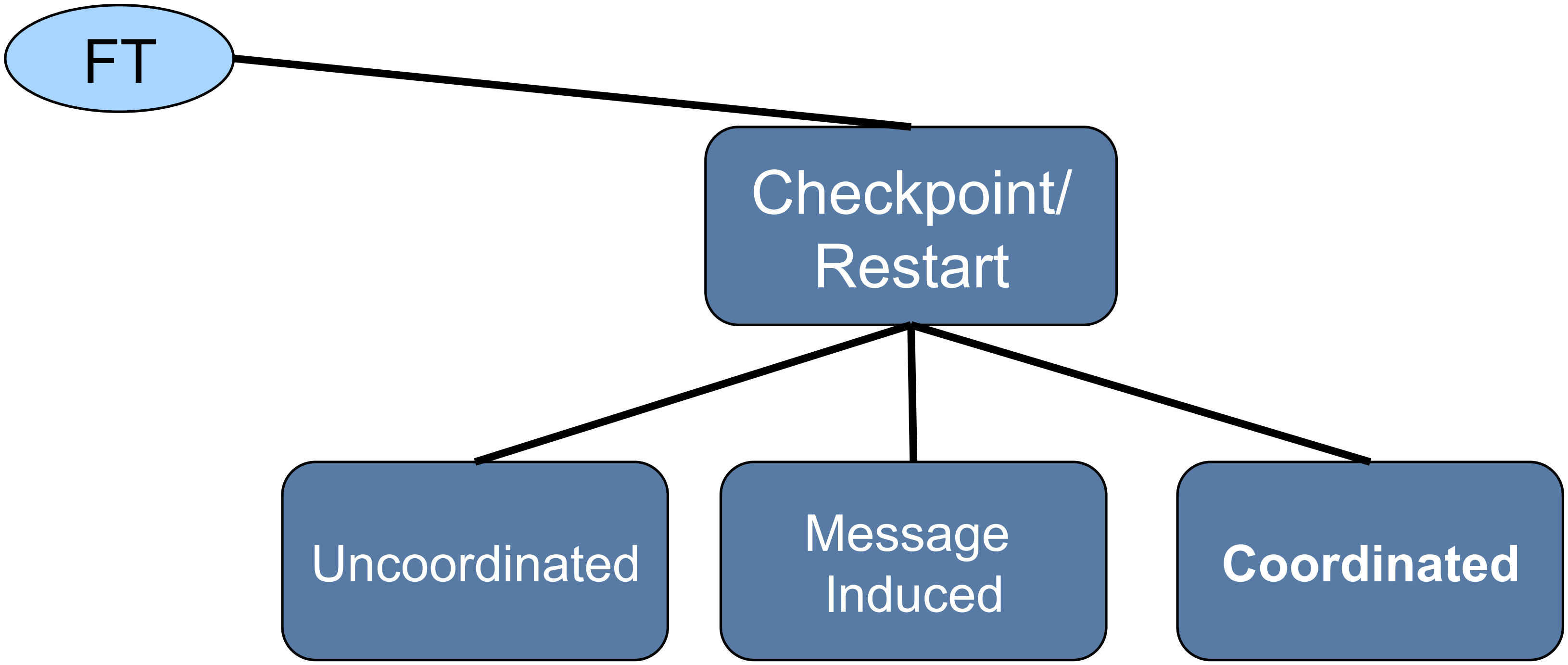
FT

C/R

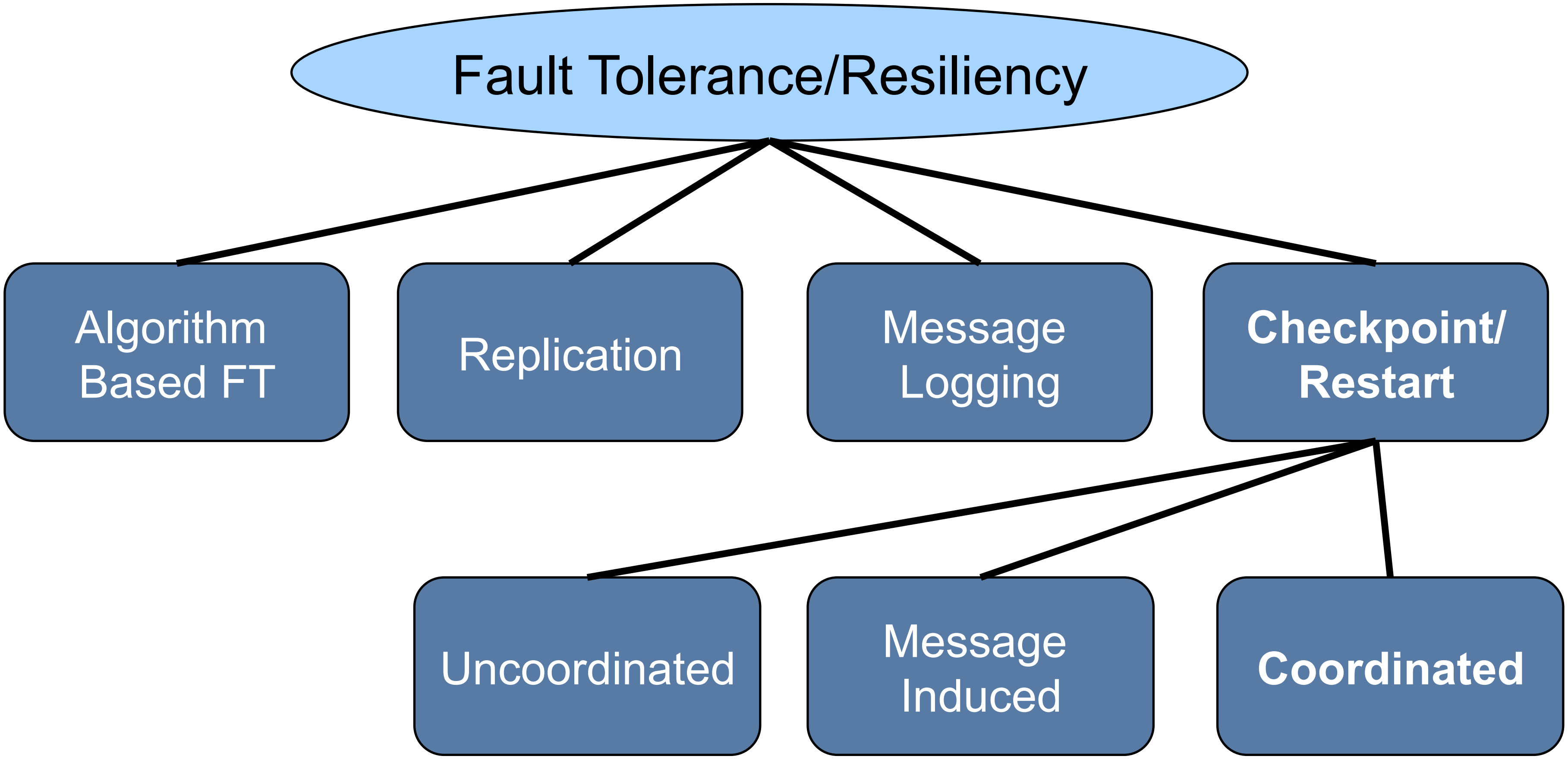
Coordinated

High Level Goals

- Deliver usable features to end users
  - Don’t publish and run
- Extensible C/R research infrastructure
  - Focused development areas
  - Apples-to-apples comparisons
  - Opportunities for public release & support



Fault Tolerance, Debugging, Process Migration



Fault Tolerance/Resiliency

Algorithm Based FT

Replication

Message Logging

Checkpoint/ Restart

Uncoordinated

Message Induced

Coordinated

Fault Tolerance, Debugging, Process Migration

Questions

Joshua Hursey
jjhursey@open-mpi.org

osl.iu.edu/research/ft/

www.cs.indiana.edu/~jjhursey/

OPEN MPI

INDIANA UNIVERSITY
PERVASIVE TECHNOLOGY INSTITUTE